UNIVERSITÀ
DI TRENTO
Department of
Information Engineering and Computer Science

Doctoral Programme in
Information Engineering and Computer Science

TOWARDS GENDER-INCLUSIVE MACHINE TRANSLATION

Andrea Piergentili

Advisor
Matteo Negri
Fondazione Bruno Kessler

Co-Advisor
Luisa Bentivogli
Fondazione Bruno Kessler

March 2026

Acknowledgments

A PhD is about not being enough, most of the time: not knowing enough, not being experienced enough, not being motivated enough. It is the people around you who bridge those gaps by sharing their knowledge, their experience, their motivation, their time, and their joy. It is only right, then, that this thesis opens by acknowledging them.

First, I wish to thank my supervisors, Luisa and Matteo, for choosing me for this PhD, which I did not take for granted then and do not take for granted now. I also thank them for everything they taught me over these three years. Through their guidance, I learned some of the most essential elements of research: rigor, ambition, and clarity. More than anything, they pushed me to grow, both as a researcher and as a person, and that was ultimately what I was looking for. No less importantly, I wish to thank Beatrice, whose contribution to this journey deserves its own mention. She gave me her time without reservation, her patience without condition, and a collaboration that shaped this work in ways I will always be grateful for.

I thank my lab colleagues, who made the day-to-day of this PhD what it was. Marco and Sara were my point of reference throughout these years whenever I needed guidance, whether on the more technical aspects of our work or simply on how to navigate the many challenges that come with a PhD. I thank Dennis, who proved to be a generous collaborator and a valuable co-author, and Lina, who was there for me through the research and everything around it, as I hope I was there for her. I also wish to mention Mauro, Roldano, Hang, and Dhairya, who were part of this journey and contributed to making the lab a place worth coming to.

A special place in these acknowledgments belongs to the friends I made along the way, who were, without question, the biggest surprise of this PhD. I did not move to Trento expecting to find some of my favorite people here, yet that is exactly what happened: fellow PhD students from other labs, each navigating their own journey, who turned into some of the most meaningful friendships I have ever had. In somewhat of a chronological order, Daniel, Penzuccio, Helena, Goli, Nick, Daniela, Francesco, Claudia, Pacio, Giacomo, and Camilla: when life gave me pickles, you brought the rest of the Big Mac.

I thank my parents for their unconditional support and for always being there for me. The fact that they never understood a single thing about what I was doing never stopped them from being proud of me. If anything, it made us more alike than they might think.

Last, and most importantly, I wish to thank Marina. She has been there for every part of this, cheering for me when things went well and keeping me alive when things were tough. She gave me motivation when I had run out of my own, and perspective when it made all the difference. A PhD takes a lot out of a person, and by extension out of the people closest to them. That she never wavered through all of it is something for which I remain deeply grateful. She made this possible.

To conclude, to everyone mentioned above: these acknowledgments represent my best effort at thanking a group of extraordinary people, and should not be interpreted as exhaustive of the gratitude and appreciation I feel towards you. Doing so requires further groundwork and dedicated discussion, which cannot fit in this (already too long) thesis. I happily leave such efforts to future work.

Abstract

Machine translation systems have become essential tools for cross-lingual communication, yet they systematically encode and perpetuate gender bias. When translating into grammatical gender languages such as Italian, Spanish, and German, these systems default to masculine forms for gender-ambiguous referents, reinforce stereotypical associations between gender and social roles, and fail to represent non-binary identities. Such biases cause symbolic and practical harm, shaping perceptions and potentially discriminating against individuals whose gender is misrepresented or erased. This thesis proposes gender-inclusive machine translation as a principled response to these challenges. Rather than focusing on binary gender bias correction, it establishes a comprehensive framework for translation that avoids undue gender marking when gender information is unavailable and accommodates all gender identities. The investigation spans five interconnected research questions, progressing from conceptual foundations through evaluation infrastructure to practical generation and deployment.

The thesis makes contributions across multiple dimensions. At the conceptual level, it investigates gender-inclusive translation across two complementary directions: conservative approaches relying on standardized linguistic resources for gender neutralization, and innovative approaches for explicit non-binary representation. It then formally defines gender-neutral translation along with desiderata guiding its application. For evaluation, it presents three benchmarks: GeNTE for English-to-Italian gender-neutral translation, its multilingual extension mGeNTE covering German, Spanish, and Greek, and Neo-GATE for neomorpheme-based English-to-Italian translation. These resources are complemented by dedicated evaluation methods, including a classifier-based approach and an LLM-as-a-Judge framework that generalizes across languages without task-specific training. Finally, a collaboration with an Italian e-learning company grounds the research in a real-world setting, yielding technical insights on integrating gender-neutral rewriting into content production workflows and stakeholder perspectives that inform design principles for deployment, emphasizing user control, explainability, and compatibility with existing authoring tools. The outcomes of the research presented in this thesis demonstrate that gender-inclusive machine translation is socially relevant, linguistically complex, and technically feasible, but poses distinctive challenges that current systems do not fully address. The resources, methods, and insights established in this thesis provide a foundation for continued progress toward systems that serve all users equitably.

Keywords

Machine Learning, Natural Language Processing, Machine Translation, Gender, Fairness

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Machine Translation (MT) has become a cornerstone technology of our interconnected world, breaking down language barriers and facilitating communication across linguistic communities. From professional translation workflows to everyday interactions with multilingual content, MT systems have woven themselves into the fabric of digital communication. As these systems increasingly mediate human expression across languages, a growing awareness has emerged that they can encode social inequalities and negatively affect perceptions of social groups and identities [439, 51, 502, 428]. They contribute to shaping how ideas, identities, and cultural nuances are represented in the target language [453, 202].

This mediating power of MT systems brings with it significant responsibility. Language is not a neutral conduit for information transfer; it encodes and reiterates social categories, cultural assumptions, and identity markers. Among these, gender represents a particularly salient dimension, one that is deeply intertwined with language structure in many linguistic systems. When MT systems automatically assign gender in reference to human beings in the absence of gender distinctions in the source, they make choices that carry social [211, 115] and practical [481, 412] implications, potentially discriminating social groups. For instance, models often enforce harmful stereotypes by associating high-status professions with men and care-giving roles with women [441, 357], attributing adjectives related to physical appearance or emotional instability to women, while reserving terms implying rationality or leadership for men [347, 489], and systematically over-generating masculine forms [408].

The challenge of gender in MT exemplifies a broader tension in artificial intelligence (AI): how can we build systems that are both high-performing and socially responsible? This question has catalyzed growing attention to **fairness** as a fundamental pillar of trustworthy AI [422, 38, 218]. Fairness demands that AI systems treat all individuals and groups equitably, avoiding unjust bias and discrimination [104, 161, 86, 379]. This principle is enshrined in

major AI governance frameworks and has become central to contemporary AI research and policy making [134, 318].

This thesis responds to these concerns by focusing on **gender-inclusive translation** as a concrete approach to achieving fairer MT systems that move beyond binary gender considerations. Rather than defaulting to gendered forms, typically masculine ones, gender-inclusive translation seeks to produce outputs that avoid unnecessary gender marking or use dedicated forms to increase the visibility of people who do not identify within the gender binary [381], thereby preventing the propagation of unwarranted assumptions and *misgendering*.[1] This goal can be pursued through two complementary approaches. The first relies on **neutralization strategies** that rephrase text to avoid gender marking altogether using standard linguistic resources, such as epicene nouns, collective terms, or impersonal constructions (e.g., translating EN *The teacher is responsible for preparing the exam* as IT *L'insegnante ha il compito di preparare l'esame*, rather than the masculine $Il_{[M]}$ $professore_{[M]}$ *ha il compito di preparare l'esame*). The second employs **linguistic innovations** such as neomorphemes and neopronouns, which introduce novel forms that can convey neutrality while also enabling the explicit, visible representation of non-binary identities (e.g., IT *Lə professorə ha il compito di preparare l'esame*). While neutralization circumvents gender expression entirely, neomorphemes offer a direct approach that can serve both neutral reference and the visible representation of non-binary identities, depending on user needs and preferences. This thesis investigates both approaches, recognizing that they address complementary needs within the broader landscape of gender-inclusive language. These approaches are particularly relevant for translation into grammatical gender languages like Italian, Spanish, and German, where gender marking is pervasive and automatic gender assignment by MT systems can lead to discriminatory outcomes.

Despite its recognized importance, research on gender-inclusive translation remains comparatively scarce. While significant work has documented the presence and consequences of gender bias in MT, relatively less attention has been devoted to developing concrete solutions that promote gender inclusivity. Substantial work has documented the presence and consequences of gender bias in MT and developed methods to improve accuracy across masculine and feminine translations, however these efforts have largely remained within the binary framework [407]. Comparatively little attention has been devoted to translation that moves beyond the binary. This gap is both conceptual and practical: the field lacks systematic frameworks for understanding gender-inclusive translation as a task, dedicated resources for evaluating

[1] Misgendering refers to using incorrect pronouns, names, or gendered language to refer to someone, and has been documented to have significant negative impacts on mental health and well-being [306, 221].

system performance, and effective methods for generating inclusive outputs.[2]

This thesis addresses this gap by focusing on multiple interconnected challenges: understanding and defining inclusive language across different linguistic systems and approaches, creating linguistic resources to support research, developing methods to generate and evaluate inclusive translations, and understanding how different systems from traditional neural MT models to large language models (LLMs) can be leveraged to produce inclusive outputs.

Addressing these challenges requires not only technical innovation but also careful attention to sociolinguistic considerations and the preferences of diverse communities [502, 260]. The twofold goal is therefore to build systems that are not only accurate in their output, but also inclusive in their treatment of gender diversity. By developing resources, methodologies, and evaluation frameworks for gender-inclusive translation, this thesis contributes to the broader effort of making language technologies fairer and more equitable. The remainder of this Chapter outlines the specific motivations driving this research (§1.1), the research questions it addresses (§1.2), the structure of this manuscript (§1.3), and the contributions made through this doctoral work (§1.4).

1.1 Motivations

Gender bias in MT is a well-documented phenomenon with tangible social consequences [312, 473, 407]. It manifests through interconnected mechanisms. First, MT systems exhibit a systematic preference for **masculine defaults**: when translating from English into grammatical gender languages like Italian, they systematically assign masculine forms to gender-ambiguous referents (e.g., "The student" $\rightarrow$ "Lo studente"$_{[M]}$) [441, 408]. Second, systems encode **stereotypical associations**, linking certain professions and roles to specific genders (e.g., nurses to feminine, doctors to masculine) and producing outputs that reinforce these biases [409, 465]. Moreover, the reliance on binary gender frameworks leads to the **erasure of non-binary identities**, as individuals who do not conform to the masculine/feminine dichotomy are either misgendered or entirely omitted from representation [419, 115, 85].

These biases are not merely technical shortcomings, they reinforce societal inequalities [58, 379]. Language shapes perception and perpetuates discrimination [439, 495, 238]: psycholinguistic research has shown that masculine generics evoke predominantly male mental imagery rather than achieving true neutrality [426, 61, 187], thus rendering women and non-

[2]Challenges surrounding gender-inclusive translation extend beyond technical and methodological aspects. In many languages, the linguistic means for expressing gender inclusivity remain limited or contested, and no consensus has emerged on which approaches should be standardized or widely adopted [447, 121, 345]. These factors shape the broader context in which any technical solution must operate, though they fall outside the scope of this thesis.

binary individuals cognitively marginalized or invisible [34]. Such representational unfairness carries practical consequences as well, as reported in recent human-centered studies [412].

Gender-inclusive language and translation offer a principled response to these challenges [415, 480]. This approach can be framed as a *de-gendering* strategy, distinct from *de-biasing* efforts that aim to reduce systematic gender skew, typically by improving accuracy across masculine and feminine outputs, without questioning the binary framework itself [408, 405]. While de-biasing addresses how gender is assigned, it offers no solution when the source text provides no gender information to begin with. Gender-inclusive translation addresses precisely this scenario: by using inclusive forms, it avoids undue gendering unless information about the referents' gender is explicitly supplied, while also embracing all gender identities.

Such forms can prevent misgendering individuals, avoid the propagation of masculine generics, and refrain from reinforcing stereotypical associations [445]. Gender-inclusive strategies span from more conservative gender-*neutralization* approaches (e.g., "chairperson" instead of "chairman") [157], to more innovative neologistic devices such as neopronouns (e.g., EN *ze/zir* instead of *he/she/him/his/her*) and neomorphemes (e.g., IT *-ə/-ɜ* instead of the masculine morphemes *-o/-i* and the feminine ones *-a/-e*) [261, 94]. By integrating these strategies into MT, systems can move beyond merely reflecting biased patterns in training data toward actively supporting fairer and more equitable communication.

The rising demand for inclusive language further motivates this research [453, 487]. Gender-inclusive practices are increasingly adopted by institutions [203] such as the European Parliament [135] as well as through updated style guidelines in academia [27] and in the industry [467, 309]. This convergence of documented bias, demonstrated harm, and societal demand creates an imperative for developing MT systems capable of producing gender-inclusive outputs.

1.2 Research Questions

The overarching goal of this thesis is to establish a comprehensive framework for gender-inclusive MT that addresses theoretical, methodological, and practical dimensions of the problem. To achieve this goal, the research is organized around five interconnected research questions that progressively build from conceptual foundations to real-world applications:

- **RQ1. How can a gender-inclusive paradigm for MT be designed to address gender bias beyond the binary?**

  This foundational question addresses the conceptual framework underlying gender-inclusive translation. Prior work on gender bias in MT has predominantly focused on

binary gender distinctions [189, 441, 357] or on controlling which gender to generate within the masculine/feminine binary [476, 438, 406, 463]. This question puts our idea of fairness beyond this dichotomy and aims to propose a new paradigm for a broader form of inclusion. Answering this question requires gathering and synthesizing insights from sociolinguistic research, institutional guidance, and technical constraints of translation systems.

- **RQ2. How can gender-inclusive translation be systematically represented and benchmarked?**

  Moving from the theoretical scaffolding to empirical grounding, RQ2 addresses the challenge of creating evaluation resources for gender-inclusive translation. Existing gender bias benchmarks have relied primarily on synthetic templates or contrastive pairs designed to test binary gender control [441, 53, 366], but these approaches are insufficient for evaluating gender-inclusive translation. Instead, we need naturally-occurring parallel data where neutralization strategies emerge organically, reflecting the variety in the linguistic solutions employed by human translators. However, a system that indiscriminately neutralizes all gender markers risks being perceived as imposing inclusive language rather than enabling it. Such benchmarks must therefore represent both scenarios requiring neutralization and contexts where gender should be preserved, enabling fine-grained assessment of systems' decision-making capabilities. This question explores how to construct natural, high-quality benchmarks that capture this complexity across multiple language pairs.

- **RQ3. How can gender-inclusive translation be automatically evaluated?**

  Standard MT evaluation metrics cannot effectively evaluate gender-inclusive translation. RQ3 investigates approaches to automatic evaluation that can handle the unique challenges of gender-inclusive translation, namely the high variability in valid neutral formulations, the need to distinguish between appropriate and inappropriate gendered outputs, and the requirement to scale across languages without language-specific fine-tuning. This question investigates whether gender-inclusive translation requires a departure from traditional reference-based evaluation paradigms, and what alternative approaches might better assess different forms of inclusive language depending on their distinctive features.

- **RQ4. How can gender-inclusive translation be automatically generated?**

  Current MT systems cannot perform gender-inclusive translation. With evaluation infrastructure in place, RQ4 addresses the core generation challenge: how can we build

systems capable of generating gender-inclusive translations? Addressing this question requires exploring the suitability of current state-of-the-art systems, generation strategies that leverage the capabilities of LLMs, and fine-tuning methods that adapt models specifically for inclusive translation. RQ4 investigates whether current generation paradigms can recognize when neutrality is appropriate and actually produce accurate, fluent, and natural inclusive translations across diverse contexts.

- **RQ5. What requirements and perspectives shape the deployment of gender-inclusive language in professional settings?**

  The final research question bridges the gap between academic research and practical deployment. RQ5 examines how gender-inclusive language is conceptualized and implemented in a real production setting. Focusing on monolingual gender-neutral rewriting, the thesis investigates how different stakeholders (project managers, developers, content designers) understand inclusivity, as well as their perspectives and requirements for the implementation and usage of automatic systems for inclusive language. The investigation reveals both the practical constraints that shape deployment and the real-world needs that should inform research priorities.

Collectively, these questions articulate a comprehensive research agenda that spans from the abstract definition of gender-inclusive translation (RQ1) through its operationalization in data (RQ2) and evaluation (RQ3), to its practical realization through automatic generation (RQ4), and finally to its embedding in real-world applications (RQ5). Each question builds upon insights from the previous ones, while also contributing independently to the broader field of fairness in language technologies.

1.3 Thesis Structure

This thesis is organized to progressively address each research question, building from theoretical foundations through empirical resources and evaluation methodologies to practical generation approaches.

Chapter 2 provides the technical and conceptual foundations necessary for the research presented in the following Chapters. It introduces the architectures underlying modern MT, from encoder-decoder models to decoder-only LLMs, and examines how gender functions in language from linguistic, sociolinguistic, and psycholinguistic perspectives. The Chapter concludes by situating this work within the broader landscape of research on gender bias and inclusivity in natural language processing (NLP).

Chapter 3 addresses **RQ1** by establishing the conceptual framework for gender-inclusive translation. It introduces the distinction between *conservative* strategies (leveraging standard linguistic resources to avoid gendered forms) and *innovative* strategies (employing neologistic devices to explicitly represent non-binary identities), and formally defines gender-neutral translation (GNT) along with desiderata guiding its application.

Chapter 4 tackles **RQ2** by discussing the development of evaluation resources for gender-inclusive translation. The Chapter presents a suite of benchmarks, GeNTE, mGeNTE, and Neo-GATE, that enable systematic evaluation of both conservative and innovative inclusive strategies across multiple language pairs.

Chapter 5 responds to **RQ3** by investigating automatic evaluation methods for gender-inclusive translation. Starting from the limitations of standard MT metrics, it develops and tests reference-free approaches ranging from a classifier-based method to LLM-as-a-Judge frameworks.

Chapter 6 addresses **RQ4** by exploring approaches to automatically generate gender-inclusive translations. It investigates the capabilities of MT systems and LLMs in zero-shot and few-shot conditions, examining prompting strategies for both gender-neutral and neomorpheme-based translation.

Chapter 7 engages with **RQ5** by examining gender-inclusive language in a real-world scenario. It explores fine-tuning approaches for gender-neutral rewriting (GNR) in Italian and investigates how industry professionals conceptualize inclusive language and define their requirements for automatic systems.

Finally, Chapter 8 discusses broader implications for fairness in language technologies, acknowledges limitations and ethical considerations, and outlines directions for future research.

1.4 Thesis Contributions

This Section presents the principal scientific contributions developed during my PhD, organized according to their nature and scope. The contributions are grounded in the following core publications, which are first listed as P1 through P8 and then mapped to the research questions presented in §1.2.

P1: **A. PIERGENTILI***, D. FUCCI*, B. SAVOLDI, L. BENTIVOGLI, M. NEGRI. 2023. *Gender Neutralization for an Inclusive Machine Translation: from Theoretical Foundations to Open Challenges*. In Proceedings of the First Workshop on Gender-Inclusive Translation Technologies.
My role: I contributed equally as a primary contributor (*) to all aspects of the work.

P2: **A. PIERGENTILI***, B. SAVOLDI*, D. FUCCI, M. NEGRI, L. BENTIVOGLI. 2023. *Hi Guys or Hi Folks? Benchmarking Gender-Neutral Machine Translation with the GeNTE Corpus*. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
I contributed equally as a primary contributor (*) to all aspects of the work.

P3: B. SAVOLDI, **A. PIERGENTILI**, D. FUCCI, M. NEGRI, L. BENTIVOGLI. 2024. *A Prompt Response to the Demand for Automatic Gender-Neutral Translation*. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics.
My role: I designed the prompting approaches, executed the experiments, and participated in the manual evaluation of the outputs.

P4: **A. PIERGENTILI**, B. SAVOLDI, M. NEGRI, L. BENTIVOGLI. 2024. *Enhancing Gender-Inclusive Machine Translation with Neomorphemes and Large Language Models*. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation.
My role: I took primary responsibility for all aspects of this work.

P5: S. FRENDA*, **A. PIERGENTILI***, B. SAVOLDI, M. MADEDDU, M. ROSOLA, S. CASOLA, C. FERRANDO, V. PATTI, M. NEGRI, L. BENTIVOGLI. 2024. *GFG – Gender-Fair Generation: A CALAMITA Challenge*. In Proceedings of the Tenth Italian Conference on Computational Linguistics.
My role: I contributed equally as a primary contributor (*) to all aspects of the work.

P6: **A. PIERGENTILI**, B. SAVOLDI, M. NEGRI, L. BENTIVOGLI. 2025. *An LLM-as-a-judge Approach for Scalable Gender-Neutral Translation Evaluation*. In Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies.
My role: I took primary responsibility for all aspects of this work.

P7: **A. PIERGENTILI**, B. SAVOLDI, M. NEGRI, L. BENTIVOGLI. 2025. *Gender-Neutral Rewriting in Italian: Models, Approaches, and Trade-offs*. In Proceedings of the Eleventh Italian Conference on Computational Linguistics.
My role: I took primary responsibility for all aspects of this work.

P8: B. SAVOLDI, G. ATTANASIO, E. CUPIN, E. GKOVEDAROU, J. HACKENBUCHNER, A. LAUSCHER, M. NEGRI, **A. PIERGENTILI**, M. THIND, L. BENTIVOGLI. 2025. *Mind the Inclusivity Gap: Multilingual Gender-Neutral Translation Evaluation with mGeNTE*. In Proceedings of the 2025 Conference on Empirical Methods in Natural

Language Processing.

My role: I contributed to refining the benchmark, designing the experiments, and executing the GNT generation and evaluation.

All the data, models, and code created during my PhD are publicly released.

Theoretical Contributions. This thesis establishes the conceptual foundations for gender-inclusive translation (**RQ1**). The first contribution, presented in P1, is the formal definition of GNT as the task of translating without marking the gender of human referents when such information cannot be reliably inferred from the source. Through systematic analysis of 30 institutional guidelines for inclusive language in English and Italian, we introduce three desiderata to guide when and how neutralization should be applied, distinguishing GNT from approaches that merely control binary gender assignment. A taxonomy of neutralization strategies is also provided, mapping the landscape of linguistic solutions available for GNT.

The theoretical framework is further extended to encompass *innovative* inclusive strategies, particularly neomorphemes, novel characters that replace gendered morphemes to convey gender-neutrality or explicitly refer to non-binary identities (e.g., IT: *scienziatə* instead of *scienziato/scienziata*, EN: *scientist*). P4 develops this extension, complementing the conservative neutralization approach with mechanisms for explicit non-binary representation, completing the spectrum of gender-inclusive translation strategies.

Resource Contributions. This thesis develops a suite of benchmarks for gender-inclusive translation across languages and approaches (**RQ2**).

GeNTE is the first natural benchmark for evaluating GNT from English into Italian. Unlike prior synthetic resources focused on binary gender, GeNTE is constructed from naturally occurring and professionally edited data, and features both scenarios requiring neutralization and contexts where gendered forms should be preserved. P2 details the corpus construction and annotation methodology, which enables comprehensive assessment of systems' processing of gender in the context of the source and their ability to generate gender-neutral outputs.

mGeNTE, introduced in P8, extends the GeNTE benchmark by adding German, Spanish, and Greek as target languages, enabling cross-linguistic investigation of GNT. This multilingual resource reveals how neutralization challenges manifest differently across grammatical gender systems and supports systematic comparison of models' behavior across languages.

Neo-GATE is the first benchmark designed to evaluate neomorpheme-based gender-inclusive translation from English into Italian. P4 presents the benchmark alongside its flexible annotation scheme, adaptable to different neomorpheme paradigms, thus addressing the evaluation needs of innovative inclusive strategies.

These resources have been leveraged to create the Gender-Fair Generation (GFG) Challenge at the CALAMITA initiative,[3] presented in P5 and designed to assess Italian LLMs on gender-fair language recognition and generation across detection, reformulation, and translation tasks.

Methodological Contributions. This thesis develops evaluation methods tailored to the unique challenges of gender-inclusive translation (**RQ3**).

GNT evaluation is a largely unexplored area that this thesis addresses through systematic investigation. A first analysis of standard reference-based MT metrics reveals their inadequacy for this task. String-matching metrics such as BLEU [337] rely on n-gram overlap and thus penalize outputs using valid neutralization strategies different from the reference, conflating surface form with neutralization quality. Neural metrics such as COMET [370] and BERTScore [522], trained on corpora where masculine generics predominate, tend to assign lower scores to neutral formulations. As a first solution, a classifier-based evaluation approach is proposed, using a model fine-tuned to distinguish gendered from neutral text for reference-free assessment in Italian. To overcome the need for language-specific training data, an LLM-as-a-Judge framework is also proposed for scalable, reference-free GNT evaluation that generalizes across languages (Italian, German, Spanish). The framework supports both sentence-level judgments and fine-grained phrase-level analyses through different prompting strategies. Finally, multilingual experiments provide a systematic cross-linguistic comparison and analysis of the proposed GNT evaluation methods. These evaluation methods are progressively developed across P2, P3, P6, and P8.

For neomorpheme-based translation, dedicated evaluation methods are developed in P4 to handle the novel morphological patterns and assess whether generated outputs correctly apply neomorphemes in appropriate contexts.

Empirical Contributions. This thesis presents a systematic empirical investigation of gender-inclusive translation generation (**RQ4**, **RQ5**).

Regarding generation capabilities, initial experiments demonstrate that MT systems and general-purpose LLMs fail to produce neutral outputs in zero-shot conditions, establishing a baseline for experiments with dedicated prompting strategies. Investigation of few-shot prompting strategies shows that GPT-4 can achieve approximately 65–70% of neutral translations when appropriately prompted, demonstrating ability to generalize beyond provided

[3] CALAMITA (*Challenge the Abilities of LAnguage Models in ITAlian*) is a collaborative effort to develop a benchmark for evaluating LLMs' capabilities in Italian on different tasks and aspects. See `https://clic2024.ilc.cnr.it/calamita/` [323].

examples. This exploration of prompting approaches applied to gender-inclusive translation is conducted in P3. The first systematic multilingual evaluation of open LLMs for GNT reveals that while models can often recognize when neutrality is appropriate, they struggle to consistently produce neutral outputs across English into Italian, German, Spanish, and Greek. This multilingual investigation is conducted in P8.

For neomorpheme-based translation, prompting experiments with multiple LLMs reveal varied capabilities and distinct failure patterns, with accuracy remaining insufficient for deployment despite promising results from some models. P4 presents these experiments and results analyses.

Regarding real-world deployment, systematic evaluation of GNR in Italian, a monolingual task relevant to industrial settings, compares few-shot prompting across multiple LLMs against fine-tuning approaches. Results show that fine-tuned compact models can match or exceed prompted larger open-weight LLMs. A key finding is the identification of a trade-off between optimizing for neutrality and meaning preservation: models fine-tuned on high-similarity data achieve better semantic fidelity but show reduced neutralization gains, highlighting the need for balanced data curation strategies. P7 explores these trade-offs and offers practical guidance for data curation in deployment scenarios.

Finally, alongside the core contributions presented above, I also collaborated on additional research works during my PhD. Although these works do not directly contribute to addressing this thesis's research questions and are therefore not discussed here, they have nonetheless played a significant role in shaping my expertise and informing the broader trajectory of my research. These works include:

P9: M. GAIDO, S. PAPI, M. CETTOLO, R. CATTONI, **A. PIERGENTILI**, M. NEGRI, L. BENTIVOGLI. 2024. *Automatic Subtitling and Subtitle Compression: FBK at the IWSLT 2024 Subtitling track*. In Proceedings of the 21st International Conference on Spoken Language Translation.
Contribution: This paper describes FBK's submissions to the IWSLT 2024 Subtitling track, covering subtitling and subtitle compression for English→German/Spanish. It presents a direct speech-to-subtitles model and a cascade system with open-source components, alongside an LLM-based approach for subtitle compression.
My role: I curated the prompting experiments.

P10: M. CETTOLO*, **A. PIERGENTILI***, S. PAPI, M. GAIDO, M. NEGRI, L. BENTIVOGLI. 2024. *MAGNET - MAchines GeNErating Translations: A CALAMITA Challenge*. In Proceedings of the Tenth Italian Conference on Computational Linguistics.

Contribution: This paper presents a CALAMITA challenge for evaluating LLMs' capabilities in English↔Italian MT. The challenge provides a benchmark with public and private portions to address data contamination concerns, along with baseline results. ***My role***: I contributed equally as a primary contributor (*) to all aspects of the work.

Key Points

This thesis proposes gender-inclusive MT as a principled response to gender bias in MT. It establishes theoretical foundations, evaluation resources and methods, and generation approaches that enable systems to produce outputs avoiding unnecessary gender marking.

Conceptual Foundations (RQ1)

- **Gap:** Research on gender in MT has focused on binary gender bias analysis and mitigation, lacking a theoretical paradigm for inclusive translation beyond the binary.
- **Contribution:** A conceptual framework encompassing conservative and innovative approaches to inclusive language and translation, a formal GNT definition with desiderata, and a taxonomy of neutralization strategies, connecting sociolinguistic guidance to MT implementation.

Evaluation Resources (RQ2)

- **Gap:** Existing benchmarks are synthetic, targeting only binary gender, and limited in language coverage.
- **Contributions:** Natural, expert-curated benchmarks: GeNTE (EN→IT) and mGeNTE (EN→IT/ES/DE/EL) for GNT; Neo-GATE (EN→IT) for neomorphemes-based inclusive translation.

Evaluation Methods (RQ3)

- **Gap:** Standard MT metrics fail to *(i)* reward correct GNTs and *(ii)* penalize undue gendered outputs, and are not suitable for evaluation of neomorphemes use in translation.
- **Contributions:** Reference-free evaluation with a classifier-based method for Italian and a multilingual LLM-as-a-Judge framework; dedicated metrics for neomorphemes use.

Generation (RQ4)

- **Gap:** MT systems and zero-shot LLMs fail to produce inclusive outputs.
- **Contribution:** Systematic investigation of prompting strategies for multiple approaches and languages, revealing both capabilities and persistent challenges.

Research to Practice (RQ5)

- **Gap:** Gender-inclusive language operationalization in real-world production settings needs exploration.
- **Contribution:** A collaboration with an e-learning company to implement a GNR system, including an investigation of industry professionals' perspective.

Chapter 2

Technical Foundations and Background

This Chapter introduces the key concepts and notions that form the foundation for the research contributions and the broader discussions presented in subsequent Chapters. The investigation of gender-inclusive MT sits at the intersection of two domains: the technical landscape of modern translation systems and the linguistic and social dimensions of gender expression across languages. Accordingly, the Chapter is organized into three main Sections. Section 2.1 introduces the technical foundations underlying contemporary MT. It outlines the Transformer architecture and its implementations for dedicated MT systems and general-purpose LLMs, and introduces relevant concepts such as model openness, explainability, and LLM-based evaluation. Section 2.2 examines how languages encode gender differently, the mechanisms through which linguistic practices become discriminatory, and the documented cognitive and social harms of gender-biased language. Finally, Section 2.3 surveys existing research on gender phenomena in MT, tracing the evolution from binary-focused bias documentation toward emerging work on gender-inclusive NLP, and examining the gaps that motivate this thesis.

2.1 Technical Foundations

The field of MT has undergone profound transformations over the past decades. Early approaches relied on hand-crafted linguistic rules, which were gradually supplanted by statistical methods that learned translation mappings from parallel corpora [247]. The introduction of neural[1] MT (NMT) [248] marked a further paradigm shift, enabling end-to-end learning with sequence-to-sequence models and attention mechanisms [450, 32, 248]. The new paradigm

[1]The term 'neural' refers to the use of artificial neural networks, computational models loosely inspired by biological neural systems that learn complex representations through layers of interconnected processing units [177].

quickly surpassed statistical approaches in translation quality across most language pairs [52, 76, 442]. Today, state-of-the-art systems in MT and generally in NLP are built upon the Transformer architecture [478], which has become the foundational framework for dedicated translation systems and general-purpose LLMs alike [525].

This Section provides an overview of the Transformer architecture (§2.1.1) and its two main instantiations relevant to this thesis: encoder-decoder models designed specifically for translation (§2.1.2), and decoder-only models that have emerged as powerful and versatile systems capable of performing translation alongside many other tasks (§2.1.3).

2.1.1 The Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. [478] in the MT field, fundamentally reshaped the landscape of NLP. Its key innovation, *self-attention*, allows the model to directly compute relationships between all positions in a sequence of token representations simultaneously, enabling both efficient parallelization across sequence positions and effective modeling of long-range dependencies regardless of their distance in the input.

With self-attention, each position in a sequence attends to all other positions to compute a new representation. Formally, given an input sequence of token representations $X = (x_1, ..., x_n)$, where each x_i is a vector encoding the linguistic properties of the corresponding token, the mechanism first projects each representation into three distinct vectors: *query* (Q), *key* (K), and *value* (V), using learned linear transformations:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{2.1}$$

where W^Q, W^K, and W^V are learned parameter matrices. Attention weights are then computed by estimating the relevance of each key to the query through scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

where d_k is the dimension of the key vectors; scaling by $\sqrt{d_k}$ prevents extremely large dot-product values that would push the softmax into regions with vanishing gradients. The softmax function normalizes the scores into a probability distribution, and the output for each position is computed as a weighted sum of all value vectors. This allows each token's representation to be dynamically informed by the entire sequence context.

While self-attention enables each token to incorporate information from the entire sequence, it has no inherent notion of token order, meaning that the same output would be computed

regardless of how the input tokens are arranged. To address this, *positional encodings* are added to the input representations to inject sequence position information. The Transformer further extends self-attention through *multi-head attention*, which runs h attention operations in parallel, each with its own learned projections:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2.3}$$

where each $\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$ operates on a different learned subspace, and W^O is a learned output projection matrix that combines the outputs of all heads. This allows the model to capture diverse types of relationships simultaneously. Each Transformer layer combines multi-head attention with position-wise feed-forward networks, using residual connections and layer normalization to enable stable training. Self-attention serves as the core computational mechanism in both encoder and decoder components, though with different masking strategies as detailed below.

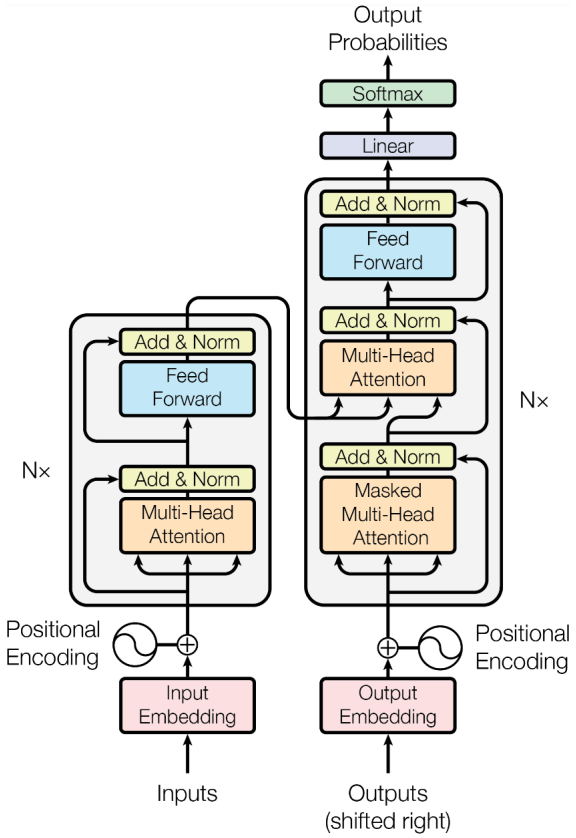

Figure 2.1: The original Transformer architecture introduced in [478].

The original Transformer architecture, represented in Figure 2.1, comprises two main components: an *encoder* that processes the input sequence, and a *decoder* that generates the

output sequence. The **encoder** applies self-attention bidirectionally, allowing each position to attend to all other positions in the input, building rich contextualized representations. The **decoder**, which generates tokens one at a time in an *autoregressive* manner, i.e., each token is conditioned on previously generated tokens. It uses causal (or masked) attention that restricts each position to attend only to preceding positions, preventing information from future tokens from leaking into the prediction of the current token. Additionally, the decoder incorporates *cross-attention* layers that allow it to attend to the encoder's output when generating each token. In cross-attention, queries are derived from the decoder's representations H_{dec}, while keys and values come from the encoder's output H_{enc}:

$$\text{CrossAttention}(H_{dec}, H_{enc}) = \text{Attention}(H_{dec}W^Q, H_{enc}W^K, H_{enc}W^V) \tag{2.4}$$

This mechanism allows each generated token to selectively attend to relevant parts of the source sequence. The encoder-decoder design, originally proposed for MT, and more generally the attention-based paradigm it introduced, became the architectural foundation for both dedicated translation systems and the broader family of language models that followed.

2.1.2 Encoder-Decoder Architectures for NMT

Standard NMT systems employ the full encoder-decoder Transformer architecture described above. These systems are trained on large *parallel* corpora, i.e., collections of source sentences paired with their translations, using maximum likelihood estimation, where the model learns to maximize the probability of the reference translation given the source. Formally, given a source sequence $\mathbf{x} = (x_1, \ldots, x_n)$ and a target sequence $\mathbf{y} = (y_1, \ldots, y_m)$, the model estimates the conditional probability as

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{m} P(y_t \mid y_{<t}, \mathbf{x}) \tag{2.5}$$

where each target token y_t is predicted based on all previous target tokens and the full source sequence, the latter accessed through cross-attention [248]. This training paradigm requires millions of parallel sentence pairs to achieve high-quality translation for a given language pair and direction [249, 108].

On the one hand, encoder-decoder NMT systems excel at the specific translation task they are trained for, achieving strong performance when sufficient data is available. Models such as MarianMT [232] and OPUS-MT [459] have demonstrated high-quality translation across numerous language pairs. More recently, large multilingual encoder-decoder systems have

pushed the boundaries of NMT coverage and scale. Meta's NLLB-200 [455, 100] supports supports over 200 languages with models scaling up to 54B parameters,[2] while Google's MADLAD-400 [254] extends language coverage to over 400 languages with models up to 10.7B parameters.[3] These systems represent the current frontier of encoder-decoder translation in terms of scale and language coverage, respectively.

On the other hand, these systems are inherently rigid: their behavior and capabilities are determined entirely by the patterns present in the training data, and adapting them to new requirements, such as producing gender-inclusive outputs or adhering to specific stylistic guidelines, requires either retraining on new data that exemplifies the desired behavior or architectural modifications to constrain the output. Since large parallel corpora reflecting gender-inclusive translation practices do not exist, standard encoder-decoder NMT systems generally perform poorly on this task (see §6.1.3) and cannot straightforwardly be trained, steered, or adapted for gender-inclusive MT. These limitations motivate the thesis's focus on LLM-based approaches and prompt-based[4] control, which can be adapted without parallel gender-inclusive data.

2.1.3 Decoder-Only Architectures, aka LLMs

An alternative paradigm to encoder-decoder models emerged with decoder-only Transformer architectures trained as language models on massive text corpora. Rather than learning to map between source and target sequences, these models are trained with a simpler objective: predicting the next token given all preceding tokens. This is known as *autoregressive* generation: the model produces tokens sequentially, with each new token conditioned on all tokens generated before it [525]. Formally, the probability of a sequence of n tokens is computed as:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_{<i}) \tag{2.6}$$

where each token x_i at position i is predicted from all preceding tokens. This objective requires only the decoder component of the Transformer, using causal attention throughout to ensure each position can only attend to previous positions, as described in §2.1.1. Early models such as GPT [362] and OPT [521] demonstrated the potential of this approach, and subsequent iterations and scaling efforts such as GPT-2 [363], GPT-3 [66], LLaMA [461], and their successors [287, 33, 508, *inter alia*] revealed that increasing both model parameters and

[2] See `https://huggingface.co/facebook/nllb-moe-54b`.

[3] See `https://huggingface.co/google/madlad400-10b-mt`.

[4] A *prompt* is the textual input provided to a language model, typically comprising task information and instructions, and possibly demonstrations that guide the model toward producing a desired output.

training data yields substantial improvements in capabilities. These large-scale decoder-only models became known as LLMs, and now represent the dominant paradigm in NLP [310, 275].

Unlike encoder-decoder NMT systems trained on parallel corpora for specific language pairs, LLMs are *pretrained* (see below) on web-scale text data spanning diverse domains, languages, and tasks [525]. This broad pretraining enables the acquisition of general linguistic knowledge and implicit representations of countless tasks encountered in the training data, while internalizing statistical structure that gives rise to knowledge-like behavior across a wide range of domains [342, 295]. Because the training corpora include multilingual text and, incidentally, translation-like content (e.g., parallel passages, bilingual documents) [57], LLMs acquire translation capabilities as a byproduct of language modeling, without explicit supervision on parallel data [25, 95, 65]. Recent research documented progressive improvement in LLMs' performance in MT, which rapidly led to a paradigm shift: LLMs are the current new state-of-the-art for MT [506, 242, 533, 114].

In-Context Learning. A defining property of LLMs is *in-context learning* (ICL): the ability to perform tasks specified through natural language instructions and task exemplars provided directly in the input context, without any update to model parameters [66, 125]. Given a task description alone (*zero-shot prompting*) or a task description accompanied by a few input-output demonstrations (*few-shot prompting*), LLMs can generalize to perform that task on new inputs. This capability represents a fundamental departure from traditional machine learning, where adapting a model to a new task typically requires collecting task-specific training data and updating model parameters [177]. With ICL, the same frozen model can be repurposed for diverse tasks simply by changing the textual prompt it receives [283].

The capacity for ICL stems from how decoder-only LLMs process information. Unlike encoder-decoder NMT systems, which maintain a distinct source-encoding stream that the decoder accesses via cross-attention (§2.1.1), decoder-only models process instructions, demonstrations, and queries as a single unified sequence. The provided instructions and examples become part of the context that informs next-token prediction, allowing the model to identify and apply the demonstrated pattern [66]. This unified processing, combined with pretraining on diverse text that implicitly contains countless task formulations, enables the generalization that ICL requires [66, 79]. Encoder-decoder NMT systems, trained on parallel corpora for a fixed source-to-target mapping, lack both the architectural flexibility and the exposure to diverse task patterns necessary for this capability.

The emergence of ICL is closely tied to model scale. Wei et al. [500] documented that certain capabilities, including effective few-shot learning, appear only in sufficiently large models, exhibiting sharp transitions rather than gradual improvements. Within the models,

this emergence correlates with the formation of 'induction heads' [326]. These are specialized attention heads within the Transformer architecture that allow the model to perform simple pattern identification and reproduction [105]. Specifically, these heads enable the model to look back at the context, recognize where the current input matches a previous example, and reproduce the completion that followed that example [532]. By chaining these operations together, the model can infer and apply the input-output relationship demonstrated in the prompt to new queries [528]. This pattern-matching capability is central to the approaches explored in this thesis: when provided with examples of gender-inclusive formulations, these mechanisms enable the model to recognize the demonstrated linguistic patterns and assess their usage in system outputs (see §5.3, where LLMs are used for evaluation) or reproduce them in new translations (see §6, where LLMs are used for generation).

The ICL capabilities of pretrained LLMs can be substantially enhanced through *instruction tuning*, which fine-tunes models on diverse tasks framed as instruction-response pairs [499, 90]. While pretraining exposes models to task-like patterns implicitly present in web text, instruction tuning provides explicit supervision on following user-specified requirements across a wide range of tasks [520].

Prompt Design. The effectiveness of ICL depends significantly on how prompts are designed [311]. In zero-shot prompting, where only instructions are provided without examples, several aspects of prompt formulation can substantially affect performance. The format and phrasing of instructions matter: minor wording changes can lead to significant differences in output quality [526]. For multilingual tasks, additional factors come into play, including whether instructions are provided in the source language, the target language, or a high-resource language like English [533]. These sensitivities can make achieving consistent performance through zero-shot prompting alone challenging.

Few-shot prompting, which augments instructions with task demonstrations, substantially mitigates these issues [264]. Research has shown that incorporating even a single exemplar significantly reduces models' sensitivity to variations in prompt formulation [182, 81]. With few-shot prompting, the selection of demonstrations becomes the primary factor influencing performance: examples that are semantically similar to the test input yield better results than randomly chosen ones [282], a finding that extends to MT [8, 515].

Beyond demonstration selection, researchers have developed strategies to elicit better reasoning from LLMs. *Chain-of-thought* (CoT) prompting instructs models to generate intermediate reasoning steps before producing a final answer, improving performance on complex tasks such as arithmetic and commonsense inference [501, 276]. In its few-shot variant, CoT prompts include examples showing not only input-output pairs but also the reasoning process

leading to each answer [501]. Remarkably, zero-shot CoT achieves similar effects through simple instruction augmentation: appending guiding instructions such as "let's think step by step" to the prompt elicits step-by-step reasoning without requiring demonstrations [250]. CoT has inspired extensions such as self-consistency decoding, which samples multiple reasoning paths and selects the most frequent answer [493].

Pre-training and Fine-Tuning. While ICL and prompting enable task adaptation without modifying model parameters, *fine-tuning* offers an alternative approach that directly updates the model's weights on task-specific data. In the predominant pre-train and fine-tune paradigm, models first acquire broad linguistic and world knowledge through large-scale pretraining, then specialize for particular tasks or domains through further training on smaller, dedicated datasets [212]. This approach can yield stronger and more consistent task performance than prompting alone, particularly when sufficient training data is available and when the target behavior diverges from patterns encountered during *pretraining*. Fine-tuning enables smaller models to achieve performance competitive with larger prompted models on specific tasks [124, 155].

Full fine-tuning updates all model parameters, which becomes computationally prohibitive as models scale to billions of parameters and risks forgetting of general capabilities [235, 252, 292]. Parameter-efficient fine-tuning methods address these limitations by updating only a small subset of parameters while keeping the base model frozen [492]. Among such techniques, Low-Rank Adaptation (LoRA) has emerged as particularly effective: it introduces additional small trainable matrices enabling adaptation with a fraction of the computational and memory cost [214, 113, 471]. In MT, LoRA has matched traditional full fine-tuning performance while reducing parameter interventions by a factor of 50 [14].

The combination of ICL, instruction tuning, and efficient fine-tuning techniques fundamentally changes the requirements for adapting models to new tasks. Rather than collecting large training datasets and retraining, users can specify desired behaviors through carefully crafted prompts; when task-specific data can be assembled, fine-tuning provides a complementary approach that embeds desired behaviors into the model. For tasks like gender-inclusive translation, where large-scale training data reflecting the desired behavior is scarce or unavailable, prompting offers an immediate pathway: instruction-tuned LLMs can potentially be guided to produce inclusive outputs through appropriate demonstrations, even though they were never explicitly trained for this task. Fine-tuning on smaller curated datasets can further reinforce these behaviors. This thesis explores both pathways: few-shot prompting is used in Chapters 5, 6, and 7, whereas fine-tuning is applied in the latter.

2.1.4 Relevant Technical Concepts

Beyond core architectures, the following complementary considerations are relevant to fully understanding how this thesis approaches gender-inclusive MT research.

Open vs Closed Models. A fundamental distinction in the current landscape concerns the degree of openness with which models are released [503]. *Closed* or *proprietary* models, such as GPT-4 [3] and Claude [22], are accessible only through APIs, without open disclosure of their architecture, training data, or model weights. While these models often achieve state-of-the-art performance across a wide range of benchmarks [328, 23, 178], their closed nature raises several concerns [60]. From a **privacy** standpoint, all data must be transmitted to external servers, which may be problematic for sensitive domains such as healthcare, finance, or legal applications [510, 529, 110]. The lack of access to model internals prevents independent verification of claims about model behavior, auditing for biases [74], and **reproducibility** of research findings, particularly as providers may update models without notice [84]. Reliance on proprietary APIs also limits **autonomy**, exposing users to arbitrary changes in pricing and terms of service [60]. Finally, **customization** possibilities are inherently limited: users cannot fine-tune these models for specialized tasks or specific linguistic requirements, and must rely on prompting alone to shape model behavior.

In contrast, *open-weight*[5] models make their trained parameters publicly available, enabling users to deploy, fine-tune, and analyze them independently [287, 226]. This openness addresses each of the concerns above. **Privacy** is preserved through local deployment, ensuring that sensitive data never leaves the user's infrastructure. Open access to model weights enables **reproducibility** [348] and scientific scrutiny [432], allowing researchers to verify claims, investigate model behavior, and build upon existing work, which is essential to responsible AI development [6, 503]. Local deployment guarantees **autonomy** from external providers and their changing policies or service conditions. Furthermore, open weights enable **customization** through fine-tuning, allowing researchers and developers to adapt models to specialized domains or specific requirements that commercial providers may not prioritize. Finally, access to model internals is a prerequisite for **explainability** techniques, discussed in the following paragraph.

Open-weight models prove essential to the experiments, analyses, and discussions presented throughout this thesis. Access to model weights enables the fine-tuning experiments discussed in this work and ensures reproducibility, as experiments can be independently veri-

[5] It is worth noting that "open-weight" does not equate to "open-source": while the former provides access to model parameters, full open-source release would additionally include training code, datasets, and documentation sufficient for complete reproducibility [503, 277].

fied, and fine-tuned models can be openly shared to support future research. Moreover, Section 7.3 discusses how they allow us to address the requirements of our industry collaboration, from data privacy to control and trustworthiness.

Explainability. Explainability is the capacity to understand and interpret how a model arrives at its outputs [380, 280, 109]. This capacity enables researchers to identify sources of bias, verify that models rely on appropriate features rather than spurious correlations, and ultimately develop targeted interventions based on these insights [46]. A prerequisite for most explainability methods is access to model internals, which, as noted above, is only possible with open-weight models. Probing classifiers, for example, involve training simple classification models on top of frozen internal representations to test for the presence of specific linguistic properties, such as morphology or syntax [96, 49]. Conversely, feature attribution methods assign importance scores to input elements to reveal which parts of the input most heavily influence the generation of a specific token [270].

In the context of gender phenomena in MT, explainability techniques have proven valuable for understanding how models encode and process gender information. For instance, explainability methods applied to instruction-tuned models have shown that they systematically overlook pronouns indicating gender in misgendered translations, a finding that informed targeted mitigation strategies [29]. Embedding analysis and attention patterns have revealed why different multilingual architectures exhibit varying levels of gender bias [485, 102]. Beyond bias analysis, explainability methods prove valuable for understanding how models utilize contextual information during generation. Feature attribution techniques can measure the contribution of different prompt components to model outputs, revealing which parts of the input context (instructions, guidelines, task exemplars) the model relies on for specific subtasks [401]. The experiments presented in §6.2 are part of a broader investigation that pairs behavioral evaluation with explainability analysis. While the explainability component is not central to this thesis, it offers complementary insights showing that recognition of when gender-inclusive translation is appropriate and actual translation generation rely on different context signals, helping explain observed performance gaps. The role of explainability in building user trust for deployed systems is discussed in §7.3.

LLM-as-a-Judge. Recently, LLMs have been successfully employed as evaluators of natural-language generation tasks [490, 285, 47] including MT, where proprietary LLMs have achieved higher correlation with human judgments than traditional automatic metrics in quality evaluation [244, 266] without requiring dedicated fine-tuning data. While human evaluation remains the gold standard, it is expensive, time-consuming, and difficult to scale. LLM-based evalua-

tion offers an effective proxy that outperforms other automatic metrics. Crucially, it enables the assessment of tasks for which no dedicated evaluation methods exist. Beyond holistic quality judgments, the *LLM-as-a-Judge* paradigm has also proven particularly useful in assessing fine-grained aspects such as fluency, accuracy, and style [154, 291], as well as to generate the error annotations required for *multidimensional quality metrics* [139, 243, 227, 536], an evaluation paradigm originally designed for human evaluators that requires pinpointed analysis and attention to context. Particularly relevant to this thesis, LLMs have also been found to be accurate evaluators of masculine and feminine references to human entities in monolingual contexts [112], suggesting their potential applicability for evaluating gender-related phenomena in translation. As we will see, unlike model-based approaches that require language-specific training data, the LLM-as-a-Judge paradigm can generalize across languages through prompting alone, making it a promising solution for scalable evaluation of gender-inclusive translation (see §5.3).

Gender-inclusive translation represents precisely a case where no dedicated automatic evaluation methods existed prior to this work. To investigate whether LLM-based evaluation could address this gap, this thesis develops and compares two reference-free approaches in Chapter 5: a classifier-based method trained on synthetic data to recognize gender-neutrality features in Italian text, and an LLM-as-a-Judge approach that assesses gender-neutrality through prompting. The classifier achieves strong performance for Italian but requires language-specific training data and cannot incorporate source-sentence information to determine whether the gender expression in the target is appropriate. The LLM-as-a-Judge approach addresses both limitations: it generalizes to new languages (Italian, Spanish, German) through prompting alone and can assess appropriateness by considering the source context, establishing it as a scalable framework for multilingual evaluation of gender-inclusive translation.

Key Points

- **Neural Machine Translation (NMT)**: Encoder-decoder Transformer systems trained on parallel corpora. They achieve strong translation quality but are rigid: adapting them to gender-inclusive translation would require retraining on data that does not exist at scale.
- **Large Language Models (LLMs)**: Decoder-only Transformers whose broad pretraining yields general-purpose capabilities, including translation. Through in-context learning (ICL), they can adapt to new tasks via prompting without parameter updates, offering a viable path to gender-inclusive translation. Instruction tuning and few-shot prompting further enhance LLMs' ability to follow task-specific

requirements.

- **Open vs Closed Models**: Open-weight models enable privacy, reproducibility, customization through fine-tuning, and access to model internals for explainability research. Closed models offer only API access, limiting these possibilities.
- **LLM-as-a-Judge**: A paradigm leveraging LLMs as evaluators of generated text, offering scalable alternatives to human evaluation without requiring dedicated training data.

2.2 Gender in Language

Gender in language reflects and constructs social categories, with different languages encoding it differently. This Section explores the relationship between gender expression and language, distinguishing between notional and grammatical gender systems (§2.2.1) and discussing how gender expression can result in discrimination and its implications (§2.2.2), ultimately introducing gender-inclusive language as a conceptual framework.

2.2.1 Gender in Language

The concept of gender is so relevant to human experience that no language lacks expressions of femaleness or maleness altogether [439]. However, languages differ substantially in how they encode gender, and this variation has significant implications for translation. Several taxonomies have been proposed for how languages embed gender within their structure [439, 188], but this thesis focuses on two widely accepted categories: *notional*[6] and *grammatical* gender languages, because they define the core translation scenario in which gender-related challenges emerge the most.

Notional gender languages, such as English, Swedish, and Danish [97, 229, 185, 222], express the gender of human referents through a limited set of linguistic elements. These include personal pronouns and possessive adjectives (e.g., *he/him/his*; *she/her/hers*), and lexically gendered forms that distinguish male and female referents (e.g., *man/woman*, *actor/actress*, *waiter/waitress*). In such languages gender marking does not extend to other parts of speech: nouns like *student*, *doctor*, or *friend* remain unmarked for gender, as do adjectives (e.g., *tall*) and verbs (e.g., *discern*). This limited gendered grammar means that English and notional

[6]While *natural gender* was the label originally used for these languages, we refer to them instead as having a *notional gender* system, to avoid confusion with terminology related to biological sex [304].

gender languages in general have faced fewer obstacles in adapting to gender-neutral forms [5], with the singular *they* already emerging as a well-established neutral pronoun endorsed by style guidelines from institutions such as the American Psychological Association [27, 222].

Grammatical gender languages like Italian, Spanish, and German are characterized by a pervasive system of morphosyntactic agreement, where gender is encoded not only in nouns but also in articles, adjectives, determiners, and, in some languages, verb forms. Consider the Italian sentence *I*/*Le bambini*/*bambine sono stati*/*state considerati*/*considerate astuti*/*astute*, (EN: *The children were considered to be astute*): here, the definite article ($I_{[M]}$ vs. $Le_{[F]}$), the noun ($bambini_{[M]}$ vs. $bambine_{[F]}$), the verb ($sono\ stati_{[M]}\ considerati_{[M]}$ vs $state_{[M]}\ sono\ considerate_{[F]}$), and the adjective ($astuti_{[M]}$ vs. $astute_{[F]}$) all carry gender inflections that must agree. This extensive marking system means that referring to a person in these languages inevitably requires selecting a gender, even when the referent's gender is unknown (e.g., *Un giorno vorrei avere* $un_{[M]}\ figlio_{[M]}$, EN: *I would like to have a child someday*) or non-binary. In these languages, all nouns belong to a gender class, but the formal gender of a word does not always correspond to the gender it conveys about its referent. The Italian word *persona* (EN: person), for instance, is grammatically feminine yet functions as conceptually gender-neutral. This thesis is concerned not with formal gender as a morphosyntactic category, but with the gender that is actually conveyed about human referents through linguistic choices. It is this conceptual dimension of gender expression that raises challenges of representation, ambiguity, and discrimination in translation. The pervasiveness of gender marking in grammatical gender languages has been linked to more visible discriminatory attitudes and greater impact of gender-biased language on social perceptions [61, 495, 222].

The notional/grammatical gender distinction adopted here deliberately simplifies a more complex typological landscape. Gender systems vary considerably in both their presence and their structure across languages [97, 127], with many languages lacking grammatical gender entirely: Finnish, Turkish, Hungarian, among others, employ invariant third-person pronouns and impose no gender agreement on nouns, adjectives, or determiners [97, 200]. At the other extreme, some languages distinguish more than two genders: German has three grammatical genders, while many Bantu languages, such as Swahili, have large noun-class systems typically treated as part of the broader domain of grammatical gender [9, 97]. Even among grammatical gender languages, variation is considerable: Hebrew and Arabic encode gender in verb morphology in ways that Italian and Spanish do not, while Slavic languages such as Russian and Polish extend agreement through complex case systems [200]. Gender systems are, moreover, not static: the increasing institutional adoption of gender-neutral pronouns in languages that previously lacked them, as with Swedish *hen*, formally incorporated into the Swedish Academy's official dictionary in 2015 [185], illustrates how social pressures for more

inclusive expression actively reshape linguistic resources.

The typological differences described above have profound consequences for translation, particularly in cross-lingual scenarios involving languages with asymmetric gender-marking systems [200, 357]. The most challenging cases arise when translating from a language with limited or no gender marking into a grammatical gender target language, creating a gender information gap: the target language demands gender specifications that the source might not provide. When translating the English sentence "The workers are tired" into Italian, for instance, a translator must make a gender choice that the source does not specify: *I lavoratori sono stanchi* (masculine) or *Le lavoratrici sono stanche* (feminine). In the absence of contextual cues, this gap must be filled arbitrarily by either human translators or automatic systems. As we will see in §2.3.1, MT systems typically resolve this ambiguity by defaulting to masculine forms or relying on stereotypical associations, with discriminatory consequences. Addressing this problem is far from straightforward: as Amrhein et al. [20] emphasize, the resources and approaches developed for English inclusive writing are not directly portable to grammatical gender languages, which require dedicated solutions that account for their pervasive morphosyntactic agreement systems. It is precisely this high-impact scenario that this thesis investigates, focusing on translation from English, by far the most widely used source language in NLP and MT research [230, 56], into grammatical gender languages such as Italian, Spanish, and German, where gender choices are forced and discriminatory patterns are most visible [495].

2.2.2 Gender and Discrimination in Language

Regardless of cross-lingual differences, linguistic practices can reinforce social discrimination when they generate a disparity in the representation of genders based on normative and stereotypical principles. Three interconnected mechanisms underlie such discrimination: androcentric normativity, gender stereotyping, and non-binary erasure [415, 231, 307].

Androcentric normativity promotes the masculine gender as the human prototype, the unmarked norm encompassing the whole human experience, thus treating women as a gendered deviation [201]. This asymmetry is deeply embedded in linguistic conventions: the masculine is treated as the default category, while the feminine is marked as a departure from that default. A typical manifestation of normativity in language is the **masculine generic**, i.e., the use of masculine forms as conceptually generic or neutral (e.g., *one must watch <u>his</u> language*) when referring to mixed-gender groups or when gender is unknown or unspecified [195]. In grammatical gender languages, this pattern extends beyond pronouns: Italian *tutti* (masculine plural) is conventionally used to address mixed-gender groups, even when

women constitute the majority (e.g., *Tutti*$_{[M]}$ *gli*$_{[M]}$ *elettori*$_{[M]}$, EN: *All constituents*) [393]. *Gender stereotyping* operates through the assumption of someone's gender based on culturally reinforced associations between social roles and gender [431]. Certain occupations become cognitively linked to one gender (e.g. nursing and teaching with women, engineering and leadership with men) [61, 188] such that hearing a role term activates gender expectations regardless of the actual referent [166, 480, 431]. These associations can shape expectations about who *should* occupy certain roles and penalize those who violate these expectations [303, 197, 208]. Finally, *non-binary erasure* occurs when linguistic systems offer only binary options for gender expression, rendering non-binary identities invisible or impossible to represent [207, 419, 72]. This form of discrimination is particularly acute in grammatical gender languages, where the binary structure of the morphological system itself, requiring every noun, article, adjective, and often verb to be marked as either masculine or feminine, excludes those who do not identify within this dichotomy. Unlike notional gender languages where neutral alternatives may exist (e.g., singular *they* in English), speakers of grammatical gender languages face a fundamental structural barrier: the grammar itself encodes a binary that leaves no sanctioned space for non-binary expression.

These discriminatory patterns are not merely linguistic conventions. They shape cognition and perception [453, 128]. Psycholinguistic research has consistently demonstrated that masculine generics fail to achieve the neutrality they purportedly convey: instead of evoking gender-neutral or balanced mental representations, they trigger predominantly male imagery [426, 61, 187]. This cognitive bias extends to large-scale language patterns: analyses of billions of words on the internet reveal systematic associations between generic human terms and male referents, effectively encoding the equation "people = men" in our collective linguistic output [34]. In turn, these biased representations influence real-world outcomes: experimental studies have shown that women are less likely to apply for positions advertised with masculine-generic language, and that gender-unfair language restricts the cognitive availability of female exemplars in professional contexts, disadvantaging women in personnel selection [444, 480, 208]. Conversely, research demonstrates that gender-fair language can help reduce gender stereotyping and mitigate these discriminatory effects [415], with evidence suggesting that gender-neutral forms in particular can increase the cognitive visibility of women and non-binary individuals [458, 138, 78].

For non-binary individuals, linguistic discrimination takes the form of systematic erasure and misgendering, i.e., using incorrect pronouns, names, or gendered language to refer to someone, which has been documented to have significant negative impacts on mental health and well-being [306, 231]. Research with non-binary populations reveals that misgendering is a pervasive experience: in a large-scale survey by Jacobsen et al. [221] 59% of non-binary

respondents reported being misgendered daily, and those who experienced more frequent misgendering showed higher levels of anxiety and depression. These findings align with broader research documenting the psychological burden of identity invalidation [534], which can lead to reduced self-esteem, social withdrawal, and in severe cases, increased risk of self-harm [460].

When these discriminatory patterns are embedded in language technologies, their impact is amplified. Machine learning systems trained on biased corpora inherit and can even magnify societal biases, propagating discriminatory representations at unprecedented scale [68, 502, 390, 161]. A useful framework distinguishes between *representational* harms, which arise when systems reinforce stereotypes or fail to represent certain groups fairly, and *allocational* harms, which occur when biased systems lead to unfair distribution of resources or opportunities [58, 470]. Both types have been widely documented in NLP: from word representations that associate professional competence with male terms [59, 430] to MT systems that default to masculine forms or rely on stereotypical associations [357, 408, 308, 489]. Such biases and harms extend to multimodal systems: vision-language models have been found to generate stereotypical completions when presented with images of men and women [391, 41], while text-to-image generation models amplify gender stereotypes when depicting occupations, traits, and activities [469, 153, 395].

These harms can also compound. Because MT is ubiquitously embedded in web applications, often invisibly to end users, people may be exposed to biased output without realizing it [297]. Moreover, representational harms frequently produce allocational ones: biased representations lead to performance disparities and uneven quality of service. Recent human-centered studies quantify these effects, showing that gender bias in MT measurably shapes user experience and perception, and imposes additional effort and economic costs as both regular users and professionals must correct biased outputs [412]. Together, these findings underscore that addressing gender bias in language technologies is not merely a matter of technical accuracy but a requirement for preventing real harm to individuals and communities.

In light of this, in this thesis we look at gender-inclusive language[7] for the avoidance of discriminatory language. Having established how languages encode gender and the mechanisms through which linguistic practices become discriminatory, we now turn to how these phenomena manifest in MT systems and the emerging efforts to address them.

[7]The label "inclusive language" covers a wide range of linguistic practices aimed at avoiding discrimination and denigration on any basis (see `https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines`). Such practices have also been given different labels, such as 'neutral' (which we consider appropriate for a subset of those practices), and 'fair' [343]. To set the object of our analysis within a larger scope of inclusivity, we hereby rely on the label *gender-inclusive language*.

Key Points

- **Languages encode gender differently**: Languages differ in how pervasively they encode gender. Notional gender languages like English mark gender only in pronouns and some lexical items, while grammatical gender languages like Italian require gender agreement across articles, nouns, adjectives, and verbs. This asymmetry creates a *gender information gap* when translating from notional to grammatical gender languages, as the target demands gender specifications absent in the source.
- **Mechanisms of linguistic discrimination**: Three interconnected patterns underlie gender-based discrimination in language: *androcentric normativity*, which treats masculine as the default human prototype (e.g., masculine generics); *gender stereotyping*, which assumes gender based on role associations; and *non-binary erasure*, which excludes identities outside the binary through systems that offer only masculine/feminine options.
- **Documented harms**: These discriminatory patterns produce measurable consequences. Psycholinguistic research shows masculine generics trigger predominantly male mental imagery, reducing women's application rates for positions and their cognitive availability in professional contexts. For non-binary individuals, misgendering causes significant psychological distress, including heightened anxiety and depression. When embedded in language technologies, these biases generate both *representational harms* (stereotype reinforcement, identity erasure) and *allocational harms* (disparity in service quality, correction costs imposed on users).

2.3 Gender Bias and Inclusivity in Machine Translation

Gender-related phenomena in language technologies are particularly salient in MT, where systems must continually make choices about how to represent human referents across languages with asymmetric gender-marking systems. Building on the technical foundations in §2.1 and the linguistic background introduced in §2.2, this Section examines how existing NLP and MT systems reproduce and amplify the discriminatory mechanisms discussed above. It first surveys evaluation resources and methods as well as mitigation strategies for gender bias in MT within a binary framework (§2.3.1), then broadens the focus to emerging work on gender-inclusive NLP (§2.3.2), situating this thesis within a shift from documenting binary bias to developing genuinely inclusive MT.

2.3.1 Gender Bias in MT

Extensive research has documented the pervasive nature of gender bias in MT systems. When confronted with the gender information gap described in §2.2.1, i.e translating from notional gender languages like English into grammatical gender languages like Italian and German, MT systems exhibit the discriminatory patterns outlined above: defaulting to masculine forms for ambiguous referents, encoding stereotypical gender-profession associations, and failing to represent non-binary identities [357, 312, 378, 408]. The following paragraphs review the evaluation resources developed to measure these biases and the mitigation strategies proposed to address them, all operating within a binary gender framework.

Gender Bias Evaluation. A substantial body of work has addressed the evaluation of gender bias in MT, producing numerous benchmarks across diverse language pairs and phenomena [408, 448] through two complementary approaches to benchmark construction: synthetic and natural. *Synthetic* benchmarks rely on template-based or artificially constructed sentences designed to isolate specific phenomena under controlled experimental conditions. This approach facilitates systematic assessment across multiple language pairs and enables precise manipulation of variables such as gender cues and stereotypicality [99]. However, synthetic constructions may sacrifice representativeness of authentic language use [62] and potentially introduce artificial patterns not reflective of real-world distributions [58, 413]. *Natural* benchmarks, by contrast, derive from authentic language sources such as encyclopedic text, parliamentary proceedings, or spoken discourse. These resources better capture the complexity and variability of actual language use, though they may be noisier and make it harder to isolate specific phenomena. Both approaches involve expert curation in their design, whether in the careful construction of templates or in the selection and annotation of naturally occurring instances. The following overview highlights some of the most influential resources in each category.

WinoMT [441], built upon the Winograd schema challenge [267], presents sentences that include two mentions of human beings with professional nouns and an ambiguous pronoun whose coreference determines the target entity's gender (e.g., *The physician told the nurse that she had been busy*), forcing MT systems to inflect the correct referent rather than default to stereotypes. It became a widely adopted resource for assessing whether MT systems correctly resolve gender when translating into eight grammatical gender languages[8] and was later expanded to cover speech translation [99]. WiBeMT [466] extended this approach to evaluate how English adjectives and verbs considered stereotypically masculine (e.g., *eminent* and

[8]Spanish, French, Italian, Russian, Ukrainian, Hebrew, Arabic, and German.

boast) or feminine (e.g., *brunette* and *dance*) determine the gender associated with ambiguous referents in translation into German. SimpleGEN [375] introduced template-generated English sentences that pair an occupation with an explicit gender cue (e.g., a kinship or a gendered noun: *That physician is a funny lady*), spanning both pro- and anti-stereotypical contexts, to test whether English ↔ German MT systems correctly inflect the occupation's gender when the evidence is unambiguous. The Multilingual HolisticBias corpus [98] extends the original HolisticBias dataset [432] to cross-lingual scenarios, pairing gender-neutral English sentences (e.g., *I'm an alcoholic*) with both masculine and feminine translations in 26 languages (e.g., Spanish: *Yo soy un*$_{[M]}$ *alcohólico*$_{[M]}$/*una*$_{[F]}$ *alcohólica*$_{[F]}$), enabling measurement of gender-related demographic bias across language pairs. GATE [366] is a manually constructed challenge set of English sentences with at least one gender-ambiguous human entity, paired with minimal masculine and feminine translations for every valid assignment (e.g., *I am tired* → Spanish: *Estoy cansada*$_{[F]}$/*cansado*$_{[M]}$), enabling contrastive evaluation of MT systems translating into Spanish, French, and Italian. It was later expanded to Turkish, Hungarian, Finnish, and Persian as GATE X-E [367].

Complementing synthetic resources, several benchmarks derive from naturally occurring parallel data. The Arabic Parallel Gender Corpus [189] offers an English–Arabic parallel dataset of sentences including first and second person mentions (e.g., *I'm glad you made it home, mom*), each annotated with labels relevant to gender expression in Arabic. Each English sentence is paired with multiple Arabic translations manually adapted to represent all possible combinations of masculine and feminine forms. MuST-SHE [53, 409, 411] is a natural benchmark for English to Italian, Spanish, French, and German test and speech translation built from TED talks. It contrasts paired feminine and masculine reference translations with their gender-swapped counterparts, distinguishing cases where speaker gender is recoverable from the audio (e.g., *I was born...* → *Sono nata*$_{[F]}$/*nato*$_{[M]}$) from those including textual gender cues (e.g., *She is a good teacher* → *elle est une bonne enseignante*$_{[F]}$/*il est un bon enseignant*$_{[M]}$), enabling targeted counterfactual evaluation. Building on MuST-SHE, gENder-IT [477] extends the English-Italian subset with word-level gender tags on all human referents and provides multiple gender-alternative Italian translations for the sentences containing gender-ambiguous entities. The BUG corpus [268] is a large-scale collection of natural English sentences drawn from Wikipedia, PubMed, and other sources, each linking a role noun to a gendered pronoun (e.g., *By checking the diary, the patient herself was able to check symptoms objectively*) and labeled as stereotypical or anti-stereotypical for evaluating gender-occupation associations in MT across eight target languages. MT-GenEval [107] builds a gender-balanced benchmark from Wikipedia with two subsets: counterfactual pairs that flip all unambiguous gender cues (e.g., *Her family* ↔ *His family*) and a contextual set

where preceding sentences disambiguate gender for neutral role nouns (e.g., *Paul intervenes and overpowers him, but he wriggles free* <sep> *The librarian is then run over by a car in front of the library and apparently killed*) to evaluate gender accuracy with contrastive references in MT from English into eight languages.[9] MiTTenS [382] combines handcrafted, synthetic, and natural texts to evaluate mistranslation from and into English across 26 languages, enabling largely automatic scoring via pronoun verification. For example, the English translation of the Spanish text *"Vino de inmediato cuando se enteró. Es una buena médica."* must use feminine pronouns.

Gender Bias Mitigation. While the resources described above enabled systematic measurement of gender bias, substantial research has also addressed mitigation strategies. These approaches can be organized into three categories corresponding to different stages of the MT pipeline: data-level interventions, which happen during pre-processing, model-level modifications directly affecting generation algorithms, and inference-time techniques taking place during post-processing [408, 473, 281]. Representative work in each category is discussed below.

Data-level interventions modify training corpora to reduce bias, primarily through counterfactual data augmentation that generates gender-swapped variants of training examples [535, 524, 301] or fine-tuning on gender-balanced datasets [101, 88]. Model-level modifications incorporate explicit gender signals, such as sentence-level speaker tags [476] or word-level target gender annotations [438] to control gender expression. Context-aware models that extend the translation window beyond individual sentences have also shown promise for gender disambiguation [48], though recent work found that such models may amplify masculine bias when context is uninformative [172]. Inference-time techniques modify the decoding process, with constrained beam search [404] and lattice rescoring [406] generating gender-diverse alternatives, though at the cost of potential quality degradation [89]. Post-processing techniques have also been developed to identify gendered morphology and re-inflect it according to user preferences [189, 10, 11].

With the rise of LLMs, research has documented persistent gender bias across model families and scales [173, 251, 473, 77, 436]. Mitigation strategies include the prompt design techniques described in §2.1.3, which can reduce bias significantly compared to straightforward prompts [399], fine-tuning on gender-inclusive data when available [40], and modular adapter-based debiasing [262]. However, the comprehensive survey by Savoldi et al. [407] concluded that there is no simple technical fix, advocating for more user-centric and multilingual approaches. Despite these advances, all approaches share a fundamental limitation:

[9] Arabic, French, German, Hindi, Italian, Portuguese, Russian, and Spanish.

following a *de-biasing* approach, they aim to improve the accuracy of binary gender assignment rather than avoid unnecessary gendering altogether. However, as noted by Saunders and Olsen [405], MT systems are frequently confronted with ambiguous inputs that convey no gender information (see §2.2.1), which calls for a *de-gendering* approach (see §3.2). Yet, the research agenda has generally overlooked this challenge. The following Section surveys the emerging body of work that begins to address this gap.

2.3.2 Gender-Inclusive NLP

Research addressing the gap identified above can be broadly organized around two approaches: the use of standard linguistic resources to achieve gender neutrality, and the adoption of neologistic devices such as neopronouns and neomorphemes. Chapter 3 develops a formal framework for these approaches in the context of MT. The following paragraphs survey research on both approaches, conducted primarily in monolingual settings, and conclude with an examination of prior work specifically targeting inclusive translation.

NLP Research on Gender-Neutral Language. In monolingual settings, research on gender-neutral language generation made notable progress, particularly for English. As we mentioned in §2.2.1, English benefits from relatively limited gender marking compared to grammatical gender languages. This linguistic simplicity facilitated the development of neutral rewriting systems. Bolukbasi et al. [59] demonstrated that models' internal word representations encode gender stereotypes, with occupation terms like *computer programmer* being closer to masculine words and *homemaker* closer to feminine ones. They proposed debiasing algorithms that modify the representation space so that gender-neutral words such as *nurse* become equidistant from gendered pairs like *she/he*, reducing stereotypical associations when generating in absence of gender information. Sun et al. [449] proposed a framework for rewriting gendered English sentences using the singular *they* pronoun and gender-neutral alternatives (e.g., replacing *policeman* with *police officer*). Their work demonstrated that neural models could learn to perform such rewrites while preserving semantic content. Similarly, Vanmassenhove, Emmery, and Shterionov [474] developed the NeuTral Rewriter, combining rule-based and neural approaches to automatically rewrite English sentences in a gender-neutral way. These systems leveraged the availability of gender-neutral linguistic devices well-established in English, where the singular *they* has gained widespread acceptance as documented in style guidelines from institutions like the American Psychological Association [27]. However, as noted in §2.2.1, the resources and approaches developed for English are not portable to grammatical gender languages [20]. Languages like Italian, Spanish, and

German lack an equivalent to the singular *they* and require extensive morphological changes to achieve neutrality, affecting not just pronouns but articles, adjectives, and verbal forms. This fundamental asymmetry meant that the approaches to English neutral rewriting could not be directly transferred to other languages. The question of how gender-inclusive language generation can be achieved for languages with pervasive grammatical gender thus remains largely open. While a dedicated system has been proposed for Italian [180], its effectiveness remains limited, as the experiments presented in § 7.2 confirm. Doyen and Todirascu [126] developed a rewriting system for French that uses collective nouns to neutralize role nouns, e.g., *soldats*$_{[M]}$ (EN: soldiers) $\rightarrow$ *armée* (EN: army). Though effective, this approach is only applicable in specific situations, namely to neutralize role nouns that have a collective counterpart. This challenge becomes particularly acute in cross-lingual settings, where the source language may lack the gender information that the target language grammatically requires. Chapter 3 addresses these issues, establishing a theoretical framework for gender-neutral MT and examining practical implications and challenges towards its implementation.

NLP Research on Innovative Inclusive Solutions. Alongside work on neutral language, a growing body of research began highlighting the importance of representing non-binary identities in NLP technologies [498]. Dev et al. [115] documented the risks and harms of gender exclusivity in language technologies and the challenges involved in non-binary representation. Their work established a framework for understanding how binary-centric systems cause representational harm through erasure and misgendering of individuals who do not conform to the masculine/feminine dichotomy. Neologistic devices such as neopronouns and neomorphemes have appeared across multiple languages as communities sought direct ways to represent non-binary identities [144, 222, 376].

In a pioneering contribution, Lauscher et al. [261] discussed the adoption of neopronouns in NLP and formulated a list of desiderata for modeling pronoun use in language technologies. Crucially, rather than treating pronoun inventories as fixed lists, they argued that NLP systems should be designed to accommodate emerging neopronoun paradigms that individuals may identify with, effectively treating the set of pronouns as an *open class*. While pronouns constitute one of the most closed word classes in traditional linguistic analysis, this perspective acknowledges that the inventory of forms in use socially is not static, and that language technologies should not assume it to be. This proved essential when dealing with the constantly evolving landscape of gender-inclusive linguistic innovations. In English, neopronouns like *ze/zir* and *xe/xem* emerged as alternatives to binary pronouns [305]. In grammatical gender languages, neomorphemes, i.e., novel characters used in place of gendered inflectional morphemes, gained traction: for instance, the schwa (ə) in Italian (e.g., *unə scienziatə* instead

of *uno scienziato*$_{[M]}$ or *una scienziata*$_{[F]}$), the *-e* suffix in Spanish (e.g., *alumne* instead of *alumno*$_{[M]}$ or *alumna*$_{[F]}$), and various graphemic innovations in French and German [192, 234, 78].

Studies highlighted the difficulty of LLMs in handling neopronouns in zero-shot settings [64, 209, 332]. Ovalle et al. [333] identified byte pair encoding (BPE) tokenization [418] as a major cause of LLMs' shortcomings, as novel neologistic forms may be split into suboptimal token sequences that interfere with fluent generation. This finding aligns with observations by Gaido et al. [159], who noted similar tokenization issues in a gender bias investigation for MT, where BPE tokenization was found to penalize feminine formulations compared to masculine alternatives. Beyond pronoun prediction, Bunzeck and Zarrieß [67] developed the SlayQA benchmark to test LLMs' ability to generalize social reasoning to contexts involving neopronouns, revealing persistent failures in tasks requiring gender-inclusive understanding.

Research also explored inclusive language in various NLP tasks, including text classification [28], co-reference resolution [71, 42], inclusive rewriting in Portuguese [479], and the impact of gender-neutral affixes on word embeddings [486]. Yet these efforts remained largely confined to monolingual settings and did not address the specific challenges of cross-lingual transfer. This thesis integrates innovative solutions into gender-inclusive MT, establishing a theoretical framework in §3.1.3, introducing dedicated evaluation resources and metrics in §4.3 and investigating generation strategies for producing translations with neomorphemes in §6.3.

Prior Work on Gender-Inclusive Translation. Within MT, investigations of inclusive translation were scarce and largely limited to proof-of-concept experiments. Gromann et al. [181] conducted a participatory study to understand the requirements and implications of gender-fair MT into German, involving stakeholders from translation practice and affected user communities. They identify key desiderata for gender-fair MT such as control, transparency, and context sensitivity, and highlight tensions between linguistic adequacy, social inclusivity, and user acceptance in real-world deployment. Subsequent work on English $\rightarrow$ German translation expanded this direction: Lardelli and Gromann [260] conducted a post-editing study where professional translators applied different gender-fair strategies (neutral rewording, inclusive characters, and neologistic approaches). Their results show that these strategies differ noticeably in terms of readability, comprehensibility, and cognitive effort, with familiar typographic solutions generally perceived as easier to process and more acceptable, while innovative approaches were often judged as more disruptive and cognitively demanding. Lardelli et al. [258] released a parallel dataset built by enriching a community-created gender-fair dictionary and sampling multi-sentence test instances from encyclopedic and parliamentary

text to enable systematic evaluation. Their analyses shows that state-of-the-art MT systems remain strongly biased toward masculine defaults and rarely generate gender-fair forms, both in isolated words and in context. Cho et al. [87] built a benchmark with template sentences to evaluate whether gender neutrality is preserved when translating from Korean into English, with both languages lacking grammatical gender, to investigate whether MT systems would use the gender-neutral devices of the target language. Their evaluation shows that commercial MT systems overwhelmingly introduce binary gendered pronouns, with genuinely neutral translations remaining rare. The analysis by Lauscher et al. [263] specifically examined how commercial MT systems handle English neopronouns, revealing systematic failures resulting in either misgendering (i.e., using masculine or feminine formulations) or low-quality outputs. This finding extends to monolingual co-reference and inference tasks [209, 452] Relatedly, Lardelli et al. [259] investigated whether commercial MT systems could serve as bilingual dictionaries for gender-fair English-to-German translation, finding that while all systems exhibited strong masculine bias, some (notably DeepL[10]) occasionally provided gender-fair alternatives for plural forms. In a broader analysis of gender bias in LLMs applied to translation, Vanmassenhove [473] reported that ChatGPT never produces gender-inclusive neomorphemes when translating ambiguous English sentences into Italian, even when the context would make such forms appropriate, though this was observed without explicitly prompting the model to generate inclusive outputs.

The most relevant prior work for translating into grammatical gender target languages was conducted by Saunders and Byrne [404], who fine-tuned English → German and English → Spanish MT models to use placeholder tags in place of determiners and inflectional morphemes (e.g., En. *the trainer* → Es. *DEF entrenadorW_END*). These placeholders could then be replaced with non-binary forms post-inference. While this is the sole work targeting the development of a gender-inclusive MT system, this approach had significant limitations: it only addressed a small subset of gendered morphology, relied on artificial placeholders rather than naturally occurring neutral forms, and was demonstrated only as a proof-of-concept without extensive evaluation or publicly available resources. What remains missing is a comprehensive treatment of gender-inclusive MT that establishes clear theoretical foundations, provides systematic evaluation resources, and develops effective generation strategies.

In summary, while research on gender-inclusive NLP has gained momentum, with work documenting the harms of binary-centric systems, proposing frameworks for modeling neopronouns, and developing initial resources for specific language pairs, this emerging body

[10] https://www.deepl.com/it/translator

of work remains fragmented. Efforts have largely focused on monolingual settings or single language pairs, evaluation resources are scarce and often limited to proof-of-concept scales, and systematic approaches to generating gender-inclusive translations are still lacking. The following Chapters address these gaps through a comprehensive research agenda spanning conceptual foundations, evaluation resources, and generation methods for gender-inclusive MT.

Key Points

- **Gender Bias in MT**: Systems translating into grammatical gender languages default to masculine forms, encode gender-profession stereotypes, and fail to represent non-binary identities.
- **Binary-Focused Research**: Existing benchmarks (e.g., WinoMT, MuST-SHE, GATE) and mitigation strategies (e.g., domain adaptation, target gender annotations) address masculine/feminine accuracy rather than inclusive alternatives.
- **Challenges for Grammatical Gender Languages**: English neutral solutions (e.g., singular *they*) do not transfer to languages requiring morphological changes across multiple word classes. Standard gender-neutral strategies must account for complex morphological agreement patterns using existing linguistic resources, while innovative forms like neomorphemes introduce additional difficulties due to suboptimal tokenization. Both approaches are hindered by data scarcity, as training corpora rarely include gender-inclusive formulations.
- **Gap in Gender-Inclusive MT**: Prior work remained limited to proof-of-concept experiments, leaving the field without conceptual foundations, evaluation resources, and generation methods for genuinely inclusive translation.

Chapter 3

Frameworks for Gender-Inclusive Machine Translation

The previous Chapter established that research on gender in MT has predominantly focused on the binary dimension of gender bias, developing methods to improve accuracy in masculine versus feminine translations while leaving the broader goal of inclusive translation for all gender identities largely unaddressed. This Chapter responds to that gap by establishing a conceptual framework for gender-inclusive MT, addressing **RQ1**: *How can a gender-inclusive paradigm for MT be designed to address gender bias beyond the binary?*

Section 3.1 examines practical approaches to gender-inclusive language, analyzing institutional guidelines for inclusive communication and developing a taxonomy of neutralization strategies, before discussing innovative non-binary linguistic resources and their implications for MT. Section 3.2 then formalizes GNT as a task for MT, articulating three desiderata that specify when and how neutralization should be applied. Finally, Section 3.3 discusses the technical challenges that GNT poses for MT systems, addressing issues of dynamic application, system constraints, and evaluation.

3.1 Understanding Gender-Inclusive Language

Gender-inclusive language is a form of *verbal hygiene* [70] by which people attempt to regulate language in conformity to certain ideals, and promote linguistic policies that reflect them [336]. Its goal is to prevent expressions that reinforce gender hierarchies or render non-binary identities invisible, promoting fairness and inclusion [415, 487, 343], in alignment with the UN Sustainable Development Goals of gender equality.[1] In grammatical gender languages

[1] See `https://sdgs.un.org/goals/goal5`

EN	I like being surrounded by my friends.
Gendered Translation	Mi piace essere circondat**o** da**i** mie**i** amic**i**.
Concise Neutralization	Mi piace avere persone amiche intorno a me. [I like to have people who are friends around me]
Verbose Neutralization	Mi piace che intorno a me siano presenti persone che considero mie amiche. [I like that there are people around me who I consider my friends]
Neomorpheme *	Mi piace essere circondat* da* mie* amic*.
Neomorpheme ə	Mi piace essere circondatə da3 mie3 amic3.

Table 3.1: Examples of EN → IT translations with no gender information in the source. The first example uses generic masculine formulations to refer to human beings (in bold), while the rest employ different gender-inclusive strategies (underlined). The second and third examples use periphrases of different verbosity, whereas the fourth and fifth ones employ different neomorpheme paradigms.

like Italian, inclusive language is both particularly challenging and increasingly urgent due to their pervasive gender systems [335, 119, 425] and the widespread use of masculine forms as default to mark generic or mixed-gender referents [186].

The efforts to make language fair and inclusive of all gender identities can be distinguished in **two main approaches** [389, 174]: the use of **gender-neutral formulations** and the introduction of **innovative non-binary linguistic resources**. Table 3.1 illustrates how these two approaches manifest in translation: in contrast with a gendered translation (first example), the *Neutralization* examples demonstrate neutralization strategies of varying verbosity relying on standard linguistic resources, while the *Neomorpheme* examples employ neomorpheme paradigms that introduce novel gender-inclusive morphological forms [447, 387].

The first approach relies on established gender-neutral devices of the standard language. While some languages already feature convenient gender-neutral resources, such as the well-established singular *they* in English [39, 18],[2] speakers of other languages, such as Italian, cannot rely on similar devices [531]. They can, however, resort to gender-neutralization strategies, such as the preference for epicene words, i.e., words that are not gender-marked and can be used regardless of the referent's gender (e.g., $docente_{[N]}$, as opposed to $maestro_{[M]}$ and $maestra_{[F]}$; EN: *teacher*). Neutralization strategies range from simple word choices to complex sentence reformulations without introducing innovative elements, thus being aligned with standardized forms and grammar. Such strategies are increasingly accepted in communication and are endorsed by institutions and organizations to embrace all gender identities [135, 203]. The second approach is instead the result of ongoing grassroots efforts, that proposed and promoted the use of innovations like neopronouns (EN *ze/zir* instead of

[2] See `https://www.merriam-webster.com/dictionary/they`.

he/she/him/his/her; FR *iel/ael* instead of *il/elle*) [376, 374], neomorphemes (ES *-e/-es* instead of *-o/-os* and *-a/-as*, as in *Hola a todes* versus $todos_{[M]}$ or $todas_{[F]}$, EN: *Hello everyone*) [156, 443, 386], and other solutions (e.g., graphemic devices such as IT -* instead of *-a/-o/-e/-i*, as in *bambin** versus $bambina_{[F.SN]}$, $bambino_{[M.SN]}$, $bambine_{[F.PL]}$, or $bambini_{[M.PL]}$, EN: *child/children*) that allow speakers to refer to individuals without resorting to binary gendered terms [192]. The acceptance of such innovative resources, however, is still highly debated and mostly restricted to informal communication channels like social media [94, 121]. Indeed, the use of gender-inclusive language has become increasingly politicized across diverse linguistic and cultural contexts, emerging as a focal point in broader societal debates over gender, identity, and language policy [344, 160]. As illustrated in Figure 3.1, media coverage reflects the contentious nature of these discussions, with headlines documenting both legislative efforts to promote inclusive forms and institutional backlash against them.

Given its alignment with standardized grammar and its broader acceptability across communicative contexts, gender-neutralization is the most viable approach for the widest range of speakers and language communities, and therefore the primary focus of this thesis. Nonetheless, recognizing the importance of innovative solutions for the explicit representation of non-binary identities, this thesis also investigates the use of neomorphemes in MT, exploring their potential and the specific challenges they pose. The remainder of this Section examines these approaches in detail: it first analyzes how institutional guidelines frame and recommend inclusive language practices (§3.1.1), then elaborates on the taxonomy of neutralization strategies shown in Table 3.2 (§3.1.2), and finally discusses innovative approaches and their implications for MT (§3.1.3).

3.1.1 Analysis of Institutional Guidelines

While MT-specific guidance for gender-inclusive language remains unavailable, resources intended for human communication offer valuable linguistic knowledge that can inform the development of inclusive translation systems. Among the most influential and accessible of these resources are the guidelines produced by renowned institutions to address gender discrimination in language. These guidelines represent what can be characterized as 'top-down' approaches to inclusive language, as they come in the form of guidance for formal communication, as opposed to the 'bottom-up' efforts emerging from grassroots movements (see §3.1.3). Although institutional guidelines address monolingual communication rather than translation, they provide a principled foundation for understanding *how* gender inclusivity is conceptualized in practice, and offer concrete insights into *what* linguistic elements are targeted and *which* strategies are recommended to achieve inclusivity.

Figure 3.1: A selection of international news headlines about gender-inclusive language across different countries and institutional contexts. The headlines reflect the politicization of inclusive language practices, from legislative proposals and bans to debates over specific linguistic devices.

To examine how these principles translate into concrete recommendations, the early stages of this PhD focused on analysing 30 guidelines published online[3] by relevant institutions, equally divided between guidelines for English and Italian (see the full list of guidelines in Appendix A). Besides prestige, comparability was prioritized: guidelines were selected from international institutions (e.g., the European Union and the United Nations) that published

[3]Retrieved through Google queries on October 28, 2022.

the same document in both languages, or from national institutions (e.g., universities and governmental bodies) that share a similar status across countries. These criteria ensured that the selected guidelines belong to the same textual genre. Table 3.2 presents examples of neutralization strategies extracted from such guidelines for English and Italian,[4] which will serve as reference for the rest of this Chapter and thesis.

Starting our analysis from ***how these guidelines interpret gender***, and gender-based discrimination consequently, clear differences emerge between the English and the Italian documents. While the former mostly go beyond the binary gender framework, the Italian guidelines tend to address women and men only. Such a difference is evident in the two versions of the European Parliament's guidelines (see documents E3, I5 in the reference list). This fundamental difference reflects different conceptualizations of discrimination (e.g., E3: "achieving equality", I5: "achieving equality between men and women"). This conceptual discrepancy is reflected in the suggested strategies to address discrimination at the linguistic level. For instance, the Italian guidelines provide extensive lists of feminine counterparts for traditionally masculine professional nouns (e.g., IT *coordinatore*$_{[M]}$ / *coordinatrice*$_{[F]}$, EN *coordinator*). They also often endorse gender specification to avoid masculine generics (e.g., EN *The professors* → IT *I professori*$_{[M]}$ *e le professoresse*$_{[F]}$). Since such recommendations remain within a binary framework, they do not conform to a broader goal of gender-inclusivity, and are hence not considered in the following discussion.

Concerning ***what linguistic elements these guidelines target***, the documents tend to focus primarily on a particular form of gender discrimination: masculine generics. Masculine generics have been historically employed in administrative and legal texts to refer to the public at large (e.g., see example B in Table 3.2, where *he* refers to the whole occupational category of *judges*, and the Italian *il docente*$_{[M]}$, 'the professor', for the full teaching body) [135, 198]. In the same vein, stereotypical associations and androcentric forms are discouraged (e.g., see example A in Table 3.2). Overall, these guidelines are mostly concerned with generic referents. As we will see in §3.2, however, there are also circumstances where avoiding gender marks is necessary to prevent misgendering specific individuals. From a linguistic standpoint, it is worth noting that English gender-inclusive strategies focus on pronouns (e.g., strategies C and E in Table 3.2), which are the main carrier of gender distinction in notional gender languages, as discussed in §2.2. The Italian guidelines, instead, prioritize nouns while overlooking adjectives, pronouns, and verbs, which are also subject to gender agreement. Although the examples in these guidelines are simple sentences within an institutional domain, effective gender-inclusive solutions should take into consideration the full range of gendered

[4] A Portuguese version of this table was created by Ferreira [141] after its publication (P1) .

A. Epicene synonyms		
EN	E5	***Chairman*** → **Chair(person)**
IT	I3	***Professore*** [Professor] → **Docente** [Teacher]
B. Pluralization (towards generic or epicene forms)		
EN	E2	A judge must certify that ***he*** has familiarized ***himself*** with... → All **judges** must certify that **they** have familiarized **themselves** with...
C. Relative and indefinite pronouns		
EN	E5	If a staff member is not satisfied..., ***he*** can ask for a rehearing. → Any staff member **who** is not satisfied... can ask for a rehearing
IT	I3	L'assicurazione... è a carico ***del fruitore*** [of the user]. → a carico di **chi fruisce** [of who uses].
D. Collective and Role nouns		
EN	✳	Please contact one of the ***waiters***. → Please contact our **staff**.
IT	I3	Il palazzo ospita gli studi ***dei professori*** [of the professors] di slavo. → Il palazzo ospita gli studi **del personale docente** [of the teaching staff] di slavo.
E. Omission		
EN	✳	A person must reside... before ***he*** may apply for permanent residence. → ...before __ applying for permanent residence.
IT	I3	Un'accurata compilazione facilita ***allo studente*** [to the student] diverse operazioni. → Un'accurata compilazione facilita __ diverse operazioni.
F. Repetition		
EN	E3	A manager may apply... if permission has been granted by ***his*** institution. → ...if permission has been granted by **that manager**'s institution.
G. Passive voice		
EN	E5	Each action officer must send ***his*** document. → Documents **must be sent**.
IT	I1	***Il*** richiedente presenta la domanda [The applicant submits the application]. → La domanda **va presentata** [The application must be submitted].
H. Imperative forms		
EN	E5	Each staff member is requested to submit ***his*** information. → Please **submit** all information.
IT	✳	***Il cittadino*** deve allegare [The citizen must attach] un documento. → **Allega** [Attach] un documento.
I. Impersonal forms		
IT	I15	***Il candidato*** decade [The candidate loses] dal diritto... → **Si decade** [One loses] dal diritto...

Table 3.2: Examples of neutralization strategies. In ***red, italic*** the generic masculine formulations; in **green, underlined** the gender-neutralizations. Column 2 provides the reference to the (E)nglish/(I)talian guidelines where each example was found (E1,2,3,..). If no example was found for a specific strategy within the guidelines, but the strategy is nonetheless applicable, we fabricated an example (indicated with ✳). If a strategy is not applicable in one language, the corresponding example was omitted.

words in grammatical gender languages.

3.1.2 A Taxonomy of Neutralization Strategies

Having examined how institutional guidelines conceptualize gender inclusivity, this Section turns to ***how gender discrimination can be avoided in language***. As previously noted, these top-down guidelines advocate for the use of neutralization strategies that conform to standardized, institutional language, rather than innovative or non-codified forms. Table 3.2 offers a systematization that maps strategies across English and Italian, except for highly language-specific solutions that cannot be transferred between the two languages.

Neutral solutions vary considerably, ranging from omissions (e.g., strategy E in Table 3.2) and simple replacements of single words with epicene or collective nouns (e.g., A, B, D), to more complex reformulations involving structural changes at the sentence level (e.g., F, G, H, I). Although elegant, noun replacement can be limiting when other gender-marked words are present, allowing only partial neutralization. Consider, for instance, the Italian sentence ***Il***$_{[M]}$ ***professore***$_{[M]}$ *è* ***tenuto***$_{[M]}$ *a rispondere* (EN *The professor must answer*): replacing the noun with an epicene alternative yields ***L'insegnante*** *è* ***tenuto***$_{[M]}$, where the masculine gender mark in the past participle remains. Moreover, the contextual nature of synonymy makes the choice of gender-neutral alternatives strictly case-specific [130]. According to the guidelines, neutralizing short segments when possible is preferable, as it produces more fluent outcomes compared to complex phrasings. This strategy is not always viable, however. Consider the Italian term *figlio*$_{[M]}$/*figlia*$_{[F]}$ (EN *child*): in the absence of epicene synonyms, neutralization would require verbose periphrases such as *persona che si è concepita o adottata* (EN *person who was conceived or adopted*).

Neutralization strategies thus emerge as complex choices that must be carefully selected and weighted to preserve communicative effectiveness and textual acceptability, including features such as fluency and style. These choices depend on various constraints, including register, length, and context of use. When adopting inclusive language, it is therefore crucial to consider the potential trade-off between neutrality and the overall acceptability of the resulting text. As illustrated by the first two examples in Table 3.1, the same source sentence can be neutralized with varying degrees of verbosity: the first employs a relatively concise periphrasis (*persone amiche intorno a me*), while the second achieves neutralization through a more complex reformulation.

The feasibility and efficacy of neutralization strategies also depend heavily on textual domain and content. Such strategies are expected to be particularly effective in certain contexts, such as administrative and institutional communication, to which most monolingual

guidelines belong. Different and less formal textual styles could present greater challenges, as the strategies discussed above might prove inapplicable or inappropriate. Consider the translation of the simple, colloquial sentence EN *I have never been there* into Italian (*Non sono mai stat<u>o</u>*$_{[M]}$/*stat<u>a</u>*$_{[F]}$ lì): none of the strategies in Table 3.2 apply here. However, compared to institutional and administrative communication, colloquial contexts tend to have greater tolerance for creative reformulations (e.g., IT *Non ho mai messo piede lì*, literally EN *I have never set foot there*). Whether MT systems should resort to such devices when straightforward neutralization strategies are not applicable is a design decision that must be considered when building inclusive systems.

These characteristics and considerations point to a fundamental tension in gender-neutral approaches. While neutralization strategies align with standardized grammar and enjoy broader institutional acceptance, they can result in verbose phrasings that are only appropriate in formal contexts [346]. The verbose formulations sometimes required by neutralization strategies, as exemplified by the third example in Table 3.1, may be appropriate for formal institutional contexts but less suited to colloquial communication, where the more natural phrasing of the gendered translation in the first example (despite its generic masculine formulations) might be pragmatically preferred by some speakers.[5] Furthermore, certain terms resist neutralization entirely in grammatical gender languages, particularly kinship terms such as *child* as mentioned saw above, or *parent*, which lacks a truly neutral Italian equivalent (*genitore* is formally masculine, *genitrice* feminine) [315]. Beyond these practical limitations, the use of circumlocutory language to avoid expressing gender has been characterized as a form of *indirect* non-binary language [30]: it conceals gender rather than explicitly representing non-binary identities. This distinction motivates the examination of innovative linguistic resources that take a *direct* and explicit approach to gender inclusivity [389].

3.1.3 Innovative Non-Binary Linguistic Resources

In contrast to the top-down institutional guidelines examined in the previous Sections, innovative gender-inclusive resources have emerged through bottom-up, grassroots efforts. As discussed in Chapter 2, neologistic devices such as neopronouns and neomorphemes have appeared across multiple languages as communities sought direct ways to convey gender-neutrality and represent non-binary identities [185, 305, 376, *inter alia*]. These devices aim to enrich language with additional resources that function as alternatives to gendered linguistic

[5] Different neutralization strategies may impact text readability to varying extents. While this represents an important research direction, the present work focuses on establishing methods for representing, producing, and evaluating gender-inclusive outputs, leaving readability assessment for future investigation (see §8.4.4).

elements, allowing for explicit inclusion of identities beyond the masculine-feminine binary [260].

The distinction between neutralization and neologistic approaches reflects different representational goals: individuals may choose neologistic devices because these forms best fit their gender identity and ideology [129, 63, 16, 376] and serve as an open statement of them, rather than using neutralization strategies that circumvent gender expression altogether [169]. This preference highlights that while neutralization strategies are valuable for avoiding unwarranted gender assumptions, they do not address the need for positive, visible representation of non-binary identities. Both approaches thus serve complementary functions within the broader landscape of gender-inclusive language.

For Italian, the primary target language in this thesis, multiple neomorpheme paradigms have been proposed and currently coexist without definitive codification [457, 447]. These proposals promote the use of specific characters in place of gendered morphemes, particularly the masculine *-o* and feminine *-a* endings (e.g., *uno scienziato*$_{[M]}$, *una scienziata*$_{[F]}$, EN: *a scientist*). The proposed neomorphemes range from letters of the Latin alphabet (e.g., *-u* $\rightarrow$ *unu scienziatu*), to typographical or graphemic symbols (e.g., the asterisk -* $\rightarrow$ *un* scienziat**) [192], to elements of the International Phonetic Alphabet [94]. Among the latter, the schwa paradigm has gained particular traction, using the IPA letter 'ə' for the singular (*unə scienziatə*) and 'ɜ' for the plural (*alcunɜ scienziatɜ*, EN: *a few scientists*) [144, 35]. Unlike most neomorphemes, which have no defined spoken realization and are therefore limited to written contexts, the Schwa paradigm benefits from an associated pronunciation: the international phonetic alphabet phoneme /ə/ for the singular form, making it applicable in both written and spoken language [35, 137, 388]. Table 3.1 illustrates how these different approaches manifest in translation, contrasting masculine generic outputs with both neutralization strategies and neomorpheme-based alternatives.

As mentioned in §3.1, the acceptance of neologistic resources remains contested. Their use is largely restricted to informal communication channels, particularly social media, and within LGBTQ+ communities [94, 387]. Critics raise concerns about accessibility, noting that the schwa character may pose challenges for screen readers and individuals with certain reading difficulties [175], and argue that it represents "a solution without a problem" [334], and it is not an effectively inclusive strategy [1]. Proponents, however, argue that language evolution naturally accommodates new communicative needs, and point to the increasing visibility of neomorphemes in public discourse [199, 487, 392]. This ongoing debate underscores that there is no single, universally accepted approach to gender-inclusive language [260].

Within this landscape, this thesis adopts a dual approach. GNT remains the primary focus, as neutralization strategies align with standardized grammar, enjoy broader institutional

endorsement, and are applicable across a wider range of communicative contexts. However, recognizing that neomorphemes address distinct representational needs and are gaining visibility [387], this thesis also investigates neomorpheme-based translation as a complementary approach. This investigation contributes to understanding how MT systems can support the full spectrum of gender-inclusive language practices, from conservative strategies that work within existing linguistic norms to innovative solutions that expand the expressive possibilities of language. Importantly, neither approach is advanced as a prescription for how language should be used. Rather, both are explored as possibilities whose adoption remains a matter of communicative choice and context (see §8.2).

Key Points

- **Gender-Inclusive Language**: A set of linguistic policies aiming to prevent expressions that reinforce gender hierarchies or render non-binary identities invisible. Two complementary approaches exist: gender-neutral formulations and innovative non-binary resources.
- **Neutralization Strategies**: Relying on standard linguistic resources (epicene synonyms, pluralization, passive voice, omission) to avoid gender marking. Institutionally endorsed and broadly applicable, but sometimes verbose and unable to explicitly represent non-binary identities.
- **Neomorphemes**: Innovative morphological forms (e.g., Italian schwa -ə) emerging from grassroots efforts to explicitly represent non-binary identities. Their acceptance remains contested and largely confined to informal contexts.
- **Contexts and Trade-Offs**: The choice and acceptability of gender-inclusive strategy depend on context: neutralization strategies are effective and welcomed in formal, institutional settings but may require verbose reformulations; neomorphemes suit informal contexts but lack broad acceptance.

3.2 Gender-Neutral Translation

With neutralization established as the primary approach to inclusive MT, the focus now turns to formalizing GNT as a concrete task. This formalization addresses a gap between the guidance available for human communicators and what MT systems require: while institutional guidelines provide strategies for achieving neutrality, they rely on human judgment and world knowledge to determine when such strategies should apply. Translating these principles into

guidance for MT systems design requires explicit criteria for neutralization as a task.

GNT is therefore defined as the task of **automatically translating from one language into another without marking the gender of human referents in the target when such marking is unnecessary or inappropriate**. Consider the English sentence *Your neighbors will thank you*: a GNT system translating into Italian should produce *Il vostro vicinato vi ringrazierà*, employing the collective noun *vicinato* (EN *neighborhood*, or *neighbors* collectively), rather than $I_{[M]}$ $vostri_{[M]}$ $vicini_{[M]}$ *vi ringrazieranno*, which uses a generic masculine formulation.[6] The goal is to prevent unwarranted gender associations while preserving the source meaning. Determining when gender marking is "unnecessary or inappropriate" depends on the information available in the source text. Here we articulate three desiderata (D) that define a *de-gendering* approach and guide when neutralization should be applied and when gendered output is warranted, with specific examples in Table 3.3.

D1: Gender should not be expressed in the output translation when it cannot be properly assumed from the source.

When the gender of a referent cannot be determined from the source text, an inclusive MT system should produce a GNT. This scenario arises frequently when translating from notional gender languages into grammatical gender ones, due to the asymmetry in gender expression discussed in §2.2.1. In such cases, neutralization prevents undue gender assumptions that may: *i)* misgender a specific referent whose gender is simply unspecified (Example 1); *ii)* exclude social groups through masculine generics, rendering women and non-binary individuals invisible (Example 2); *iii)* reinforce stereotypical associations between gender and occupation (Example 3); or *iv)* perpetuate androcentric expressions that position masculinity as the default (Example 4).

D2: Proper expressions of gender should be generated in the output translation if they are (indirectly) expressed in the source.

While D1 specifies when to neutralize, this desideratum addresses the converse: when gender information is available, the translation should reflect it. This complementary principle ensures that GNT is not misunderstood as the radical elimination of all gender marking. The goal is not to erase gender from translation but to prevent its arbitrary imposition: gender should be expressed when known, and neutralized when it would otherwise be assumed without basis. Gender can often be inferred through linguistic elements that function as *gender cues*. In English, these include third-person pronouns (*he/him/his*, *she/her/hers*), terms of address (*Mr./Mrs./Ms.*), and gender-specific nouns (*boy, lady, lord, wife*). In Example 5, the reflexive pronoun *herself* unambiguously identifies the referent as feminine, warranting a

[6]While *vicinato* is formally masculine, as a collective noun it functions as conceptually neutral.

(1)	EN	I refuse to give up on **a single student** in my class.
	IT	Mi rifiuto di lasciare indietro ***un solo*** **studente** nella mia classe.
	GNT	Mi rifiuto di lasciare indietro **qualsiasi studente**[any student] nella mia classe.
(2)	EN	**A lot of innovative teachers** began bringing comics...
	IT	***Molti*** **insegnanti** ***innovativi*** iniziarono a portare i fumetti...
	GNT	**Un gran numero di insegnanti all'avanguardia**[a large number of innovative teachers] iniziò a portare i fumetti...
(3)	EN	We train **nurses** to do it, and they use local anesthetics.
	IT	Formiamo ***le infermiere*** a farlo, e loro usano anestetici locali.
	GNT	Formiamo **il personale infermieristico**[the nursing staff] a farlo, e loro usano anestetici locali.
(4)	EN	Vehicles may only proceed **at walking pace**.
	IT	I veicoli possono procedere solo ***a passo d'uomo***[at man's pace].
	GNT	I veicoli possono procedere solo **a passo di persona**[at person's pace].
(5)	EN	Even **the founder herself** abandoned the project.
	IT	Persino ***la fondatrice stessa*** ha abbandonato il progetto.
(6)	EN	It affects one to two percent of the population, more commonly **men**.
	IT	Riguarda dall'uno al due percento della popolazione, ed è più comune ***negli uomini***.
(7)	EN	Earth was pristine before **men** appeared.
	IT	La Terra era incontaminata prima della comparsa ***degli uomini***.
	GNT	La Terra era incontaminata prima della comparsa **degli esseri umani**[of human beings].
(8)	EN	The **fishermen** were so upset about not having enough fish to catch that...
	IT	I ***pescatori*** erano così ***disperati*** per la mancanza di pesce da pescare che...
	GNT	**Le persone che pescavano**[the people who were fishing] erano così **disperate** per la mancanza di pesce da pescare che...
(9)	EN	Now when I was a **freshman** in college, I took my first biology class.
	IT	Quando ero **uno** studente al primo anno di università, seguii il mio primo corso di biologia.

Table 3.3: Examples for D1–3. We mark binary gender-marked expressions in ***red***, and in **green** those that are neutral.

feminine translation in Italian. Proper names, however, should not be treated as reliable gender cues. Names can be ambiguous across genders and cultures[7] and cannot be considered a dependable index of gender identity [400, 116, 168]. Beyond textual cues, gender may also be conveyed through non-linguistic information, such as speaker metadata provided to translators. When gender is reliably indicated through any of these means, the translation should express it accordingly.

[7] *Andrea*, for instance, is typically (but not exclusively) masculine in Italian but feminine in German.

D3: Masculine generics should not be propagated from the source language to the output translation.

The previous desiderata might suggest a straightforward distinction: neutralize when gender is absent, express it when cues are present. However, the boundary between genuine gender cues and masculine generics is not always clear. Terms like *man* and its compounds (e.g., *chairman*) can function either as gender-specific references or as generics for all humans, and this ambiguity requires careful treatment. When masculine terms genuinely refer to male individuals, gender should be preserved. In Example 6, *men* denotes the male population as a demographic category; neutralization would distort the meaning. Conversely, when such terms function as generics, referring to humanity as a whole (Example 7) or to mixed-gender categories like *fishermen* (Example 8), they should be translated with neutral forms. Propagating masculine generics from source to target perpetuates the exclusionary patterns that D1 seeks to avoid. Given the limited context within which MT systems operate, ambiguous cases arise frequently. When the generic or specific reading cannot be reliably determined, neutralization represents the safer choice: it avoids perpetuating potential masculine generics without compromising semantic content. One exception concerns first-person self-reference (Example 9): when speakers use gendered forms to describe themselves, this choice should be respected in translation, on the assumption that individuals select expressions appropriate to their own identity [17, 353]. Treating first-person gendered self-reference as intentional represents a reasonable default that avoids overriding speakers' linguistic choices.

Together, these three desiderata provide a principled framework for determining when gender should be neutralized, expressed, or carefully disambiguated. Implementing them in MT systems, however, poses significant technical challenges: from detecting gender cues and disambiguating masculine generics to generating fluent neutral formulations in grammatical gender languages. The following Section discusses these challenges in detail.

Key Points

- **Gender-Neutral Translation (GNT)**: The task of automatically translating from one language into another without marking the gender of human referents in the target when such marking is unnecessary or inappropriate.
- **Desideratum 1 (D1)**: Gender should not be expressed when it cannot be properly assumed from the source, avoiding misgendering, masculine generics, stereotypical associations, and androcentric expressions.
- **Desideratum 2 (D2)**: Gender should be expressed when reliably indicated through

linguistic cues in the source (pronouns, titles, gendered nouns) or external metadata.

- **Desideratum 3 (D3)**: Masculine generics in the source should not be propagated to the target. In ambiguous cases, neutralization is the safer choice.

3.3 Challenges and Insights for a Gender-Neutral Machine Translation

While the three desiderata articulated in Section 3.2 provide clear guiding principles for when and how neutralization should be applied, implementing them in MT systems is far from straightforward. From a formal perspective, GNT can be understood as a *constraint* on the translation process [163]: the system must produce outputs that avoid inappropriate gender marking while preserving semantic content and maintaining acceptable quality. However, the multifaceted nature of GNT makes it a particularly challenging constraint to satisfy. Gender inclusivity encompasses both *lexical* and *syntactic* dimensions, as neutralization can be realized through specific word choices (e.g., epicene nouns) or through structural reformulations (e.g., passive voice, impersonal formulations). Moreover, it involves *utility* considerations: neutral reformulations must remain fluent, coherent, and faithful to the source, characteristics not always easily guaranteed [472]. The efficacy with which neutralization is achieved can vary considerably, as alternative solutions may differ in their verbosity and semantic proximity to the source. Gender inclusivity thus combines multiple constraint types simultaneously, demonstrating a higher level of complexity than constraints that target only lexical content or syntactic structure in isolation.

Three interconnected challenges emerge as critical. First, the desiderata create a *dynamic constraint* that requires context-sensitive decisions about when to neutralize versus when to preserve or express gender (§3.3.1). Second, the fundamental absence of training data featuring high-quality and consistent GNTs hinders system development (§3.3.2). Third, the intrinsic variability of neutralization strategies complicates evaluation, as multiple structurally different solutions constitute equally valid translations (§3.3.3). The following subsections examine each challenge in detail, identifying both the universal obstacles that all MT architectures face and the architecture-specific considerations that shape potential solutions.

3.3.1 Addressing the Dynamic Nature of Gender Inclusivity

The three desiderata articulated in §3.2 create a *dynamic constraint*: before discussing *how* MT systems could produce GNTs, they need to be able to discern *when* a gender-neutral output is appropriate. D1 requires neutralization when gender cannot be properly assumed from the source, while D2 mandates preserving gender when reliably indicated through linguistic cues or external information. D3 adds further complexity by requiring systems to distinguish masculine generics from genuine masculine references. Together, these principles demand non-trivial context-sensitive decision making that goes beyond simple rule application.

The difficulty of this dynamic constraint is amplified by the gender information gap discussed in §2.2.1: the source language may not provide gender information that is required in the target. This challenge is complicated by the fact that MT systems typically operate at the sentence level, translating each sentence in isolation without access to surrounding discourse [248, 354]. Consider the English text: *He was talking with a young man. Only later I realized that this person was a professor.* When translating the second sentence independently, a system encounters no explicit gender cues: both *this person* and *professor* are gender-neutral forms in English. Yet the gender information needed to make an informed decision is located in the first sentence, where the pronoun *he* and the noun *man* establish a masculine gender for both referents. Without access to this prior context, a sentence-level system must either apply a default gendering strategy or attempt neutralization without knowing whether the referent's gender has already been established in the discourse. This context dependency poses challenges for all MT systems, though the nature of these challenges varies by architecture.

While document-level MT approaches have emerged to address discourse phenomena such as coreference and coherence [48, 290], the benefits of broader context do not automatically transfer to gender-related decisions. Basta et al. [45] demonstrated that incorporating document-level context (the previous sentence) and speaker metadata into a decoder-based NMT system yields modest improvements in binary gender accuracy (+5% on WinoMT). However, whether such contextual benefits extend to the more nuanced task of determining when to neutralize rather than simply which binary gender to assign remains an open question.

LLMs, by contrast, can process longer context windows and have demonstrated capabilities in discourse understanding across various tasks [237, 525]. In the context of MT specifically, Wang et al. [491] found that LLMs outperform both commercial MT systems and specialized document-level MT methods on discourse phenomena including coherence and cohesion, suggesting their potential for context-informed translation decisions. LLMs' ability to maintain coherence over extended passages suggests potential for making context-informed neutralization decisions. However, possessing longer context windows does not automatically

ensure appropriate neutralization decisions. These considerations are central to Chapter 6, which discusses how effectively leveraging LLMs' contextual capabilities for GNT requires careful prompt design that guides the model to attend to relevant information and apply the desiderata appropriately (§6.1). The flexibility that LLMs offer through prompt-based control provides an avenue for addressing the dynamic nature of gender inclusivity without the architectural constraints of encoder-decoder systems, though realizing this potential requires systematic investigation.

An alternative approach to context-dependent decision-making involves providing systems with explicit external information that disambiguates gender. Both encoder-decoder and LLM architectures can leverage such metadata when available. For encoder-decoder systems, gender information can be supplied through tags appended to the source input, either at the word level to indicate specific referents' gender [438] or at the sentence level to convey speaker or subject attributes [476, 45]. These approaches have proven effective for controlling binary gender assignment in translation, and similar mechanisms could in principle guide neutralization decisions. For the decoder-only LLMs, such information can be incorporated through prompts or system messages, instructing the model about when to apply neutralization based on explicitly provided context. This approach was proven effective in masculine/feminine gender control [398, 265], but inclusive translation beyond the binary remains unexplored. The limitation of these metadata-based approaches is their dependence on the availability of accurate external information, which in many real-world translation scenarios is rarely available.

The dynamic nature of gender inclusivity thus presents a fundamental challenge that all MT systems must address: determining when neutralization is due versus when gender should be preserved, or even when ambiguity in the source reflects genuine uncertainty that should not be resolved. Determining *when* to neutralize, however, is only a part of the problem. The following subsection examines the challenge of gender-neutral forms generation and the fundamental obstacles that all architectures face in acquiring this capability.

3.3.2 Constraining MT Systems Towards GNT

Beyond determining when to apply neutralization, MT systems must be capable of generating appropriate gender-neutral forms. This generation challenge is rooted in a fundamental data scarcity problem that affects all MT architectures equally: large-scale parallel corpora featuring consistent GNTs in the target language simply do not exist.

Standard parallel datasets used to train MT systems were not created with gender inclusivity as a consideration, and naturally occurring translations overwhelmingly employ default gendered forms rather than neutral alternatives [112]. For encoder-decoder NMT systems, this

data scarcity means that the patterns required for gender-neutral generation are absent from their training corpora. For LLMs, despite their exposure to massive amounts of diverse text during pretraining [65], systematic examples of GNT remain exceedingly rare in the web-scale data on which these models are trained. The absence of adequate training data calls for alternative approaches for both architectures, though possible solutions differ fundamentally in their nature and trade-offs.

For encoder-decoder NMT systems, the inflexibility imposed by training data dependence is particularly drastic. As discussed in §2.1.2, these systems learn to maximize the probability of reference translations given source inputs, and their behavior is essentially determined by the patterns present in the parallel corpora on which they were trained [450, 32, 248]. Adapting such systems to new requirements like gender-neutral output generation requires either retraining on data that exemplifies the desired behavior or implementing architectural modifications that constrain the generation process. Since the requisite training data is unavailable, research has explored various constrained generation techniques borrowed from the broader field of constrained natural language generation [204, 163, 355, 517].

Approaches to constrained generation can be organized into three main categories, each with distinct mechanisms and limitations for the GNT task. The first category involves making constraints explicit through input augmentation, where target words, lemmas, or structural specifications are appended to the source input to guide the model toward desired outputs [123, 434, 83, 322]. While this approach has proven effective for incorporating terminology constraints where specific technical terms must appear in translations, it presents significant limitations for GNT. Input augmentation techniques operate primarily at the word level and assume the availability of bilingual dictionaries that map source terms to target constraints. As mentioned above, however, for GNT such dictionaries do not currently exist, and the constraint is not simply lexical but involves morphology and syntax too. Moreover, as illustrated in §3.1.2, neutralization strategies range from simple epicene noun substitutions to extensive sentence reformulations, making it impossible to map specific constraints to any given input. The technique could potentially be applicable if comprehensive and systematic gender-neutral terminology resources were developed,[8] particularly for cases where source neutral terms can be reliably mapped to target epicene equivalents, but creating such resources remains an open challenge.

The second category encompasses decoding-time approaches that restrict the search space during generation to ensure outputs satisfy predefined constraints. Grid beam search [204] and fast constrained decoding algorithms [355] modify the standard beam search procedure

[8] Such resources would also vary significantly across languages. Even within grammatical gender languages, the parts of speech that express gender can differ (see §6.2).

to guarantee that translation hypotheses contain specific words or phrases before the search concludes. Work on gender in MT has adapted these techniques to improve gender diversity in translation, with Saunders and Byrne [404] and Saunders et al. [406] designing constrained beam search methods to generate synthetic masculine and feminine alternatives in the n-best list. However, these approaches focused on binary gender assignment rather than neutralization, and extending them to GNT faces conceptual and practical obstacles. The challenge lies in specifying which constraints to enforce when multiple valid neutralization strategies exist for the same input, as discussed in §3.1.2. Furthermore, constraining the search space can degrade translation quality, as the restriction forces the decoder toward outputs that satisfy the constraint even at the expense of unnatural, ungrammatical, or unfaithful translations [89]. The trade-off between constraint satisfaction and overall quality becomes particularly pronounced when constraints involve complex structural changes rather than simple lexical insertions.

The third category involves reranking approaches, where multiple translation hypotheses are generated and then scored according to how well they satisfy desired constraints, with the highest-scoring hypothesis selected as the final output. This strategy has been applied to various constrained MT scenarios, including integrating dubbing constraints [394] and controlling gender-specific translations [406]. For GNT, reranking could in principle score hypotheses based on the presence or absence of gendered morphology, selecting *more neutral* alternatives from the n-best list. Yet this approach presupposes a reliable method for identifying which hypotheses actually are gender-neutral and ranking them, a non-trivial evaluation challenge discussed below (§3.3.3). Moreover, research across multiple constrained MT applications has consistently shown that both constrained decoding and reranking methods risk degrading translation quality. Studies on lexically constrained decoding have documented trade-offs between constraint satisfaction and overall quality [355, 194, 89], with methods achieving high constraint accuracy resulting in lower BLEU and COMET scores [136, 518]. This consistent pattern suggests that constraining generation toward specific outputs risks selecting hypotheses that satisfy the constraint at the expense of fluency, grammaticality, or faithfulness. Moreover, reranking assumes that acceptable neutral translations appear among the generated hypotheses, which may not occur if the underlying model has not been exposed to such patterns during training.

Across all three categories, a fundamental obstacle emerges: these techniques assume that constraints can be specified precisely and unambiguously, yet GNT involves inherently variable and context-dependent choices. As Table 3.2 illustrated, the same semantic content can be neutralized through significantly different surface forms, and the appropriate strategy depends on context, register, and tolerance for verbosity. Combined with the quality trade-offs documented above, these limitations make constrained generation approaches impractical for

GNT at scale. These challenges motivated this thesis' exploration of LLM-based approaches, whose instruction-following capabilities and prompt-based adaptability offer an alternative paradigm that circumvents many of these obstacles without requiring dedicated parallel data or architectural modifications. Sections 6.1 and 6.3 investigate prompting approaches to perform GNT and translation with neomorphemes, respectively. Regardless of the generation approach adopted, however, a third fundamental challenge remains: evaluating whether the generated translations successfully achieve gender-neutrality.

3.3.3 Evaluating Gender-Neutral Outputs

Evaluating whether systems successfully produce GNTs poses fundamental challenges that affect both research and system development. The lack of dedicated test sets and metrics prevents the possibility of determining whether systems are actually making any advancements toward GNT. For this task, benchmarks should ideally comprise a range of source sentences aligned with target translations expressing either gender-marked or gender-neutral forms, depending on what the desiderata require. As a suitable starting point, such test sets could target and draw from institutional and administrative communication. This domain choice is motivated by two considerations. First, the guidelines analyzed in §3.1.1 belong to this domain, testifying for the demand for GNT in these contexts and providing established conventions for what counts as appropriate neutralization. Second, as discussed in §3.1.2, neutralization strategies are expected to be particularly effective in formal, institutional settings where generic referents are more frequent and more verbose reformulations are stylistically acceptable.

Automatic MT evaluation methods typically involve comparing system outputs with reference translations and measuring the degree of overlap between n-grams [337, 350] or the distance between the generated sentence and the reference in terms of edit operations required to make them equal [433]. More sophisticated metrics take into account not only exact matches but also stems, synonyms, and paraphrases when comparing the output with the reference [36]. Neural metrics use models to predict the similarity between the output and reference, or even directly between the source and output [370, 371]. Although metrics that do not rely solely on surface similarity may be more appropriate for evaluating gender neutrality, it may be preferable to develop accuracy-based scores that isolate the evaluation of gender neutrality from overall translation quality, to assess the two aspects separately.

One approach involves annotating gender-related expressions in the reference translation and attempting to match them in system outputs, as done in MuST-SHE [53]. In such cases, accuracy is determined through string matching between expressions in the reference and in the output. The risk of mismatch remains present, however, as automatic neutralizations may

be difficult to detect in an evaluation pipeline based on a single reference and may require extensive manual analysis to be identified [409]. Using multiple references [359] that contain different neutral realizations could alleviate this difficulty by accounting for the variability of neutralization strategies discussed in §3.1.2.

Another option would be to calculate accuracy without exploiting reference translations, as in WinoMT [441]. In WinoMT, the aim is to identify gendered translations through word alignment with the source, determine their gender through a morphological analyzer, and then check whether they correspond to the source. However, GNT presents an additional challenge: in grammatical gender languages, gender-neutral expressions may carry formal gender even when functioning as conceptually neutral. For example, Italian *la persona interessata*$_{\text{[the interested person]}}$ serves as a gender-neutral alternative to the masculine generic *l'interessato*$_{\text{[the interested one]}}$, yet it is formally feminine in grammatical gender. Morphological analysis would classify it as feminine, potentially misidentifying a successful neutralization as a gendered output. This distinction between formal and conceptual gender complicates automatic evaluation of neutrality.

Overall, effectively evaluating whether MT system outputs are gender-neutral or gender-marked presents several interconnected challenges. These challenges must be addressed to develop accurate evaluation approaches that overcome the limitations of general translation quality metrics [296, 176] and account for the intrinsic variability of gender-neutral solutions. Chapter 4 addresses the resource gap by presenting GeNTE, the first benchmark for evaluating GNT from English into Italian, along with its multilingual extension mGeNTE and the Neo-GATE benchmark for neomorpheme evaluation. As part of the exploration of evaluation methods for GNT, Chapter 5, investigates the possibility of using LLMs as evaluators capable of handling the variability inherent in gender-neutral outputs, potentially circumventing the limitations of reference-based evaluation discussed above (§5.3).

Having established the conceptual foundations for gender-inclusive translation and GNT, and identified the key challenges that MT systems face in achieving it, we now turn to the practical infrastructure required for systematic research. The following Chapter presents the evaluation resources developed to address the benchmark gap just identified, providing the foundation upon which subsequent investigations of generation methods and evaluation protocols are built.

Key Points

- **GNT as a Dynamic Constraint**: MT systems must make context-sensitive decisions about when to neutralize versus when to preserve gender, requiring access to discourse context that often extends beyond sentence boundaries.
- **Constrained Generation Limitations**: Approaches developed for encoder-decoder systems (input augmentation, constrained decoding, reranking) face obstacles for GNT due to the variability of valid neutralization strategies and the risk of quality degradation.
- **Data Scarcity**: Large-scale parallel corpora featuring consistent GNTs do not exist, limiting training-based approaches for both NMT systems and LLMs.
- **Evaluation Challenges**: Existing data and standard MT metrics are insufficient for GNT evaluation. The lack of dedicated benchmark data, the variability of valid neutral solutions and the distinction between formal and conceptual gender complicate automatic assessment.

Chapter 4

Gender-Inclusive Translation Evaluation: Data

The previous Chapter established the conceptual foundations for gender-inclusive translation, defining GNT and its guiding desiderata while identifying the technical challenges that MT systems face in achieving it. Among these challenges, the lack of dedicated evaluation resources emerged as a fundamental obstacle: without appropriate benchmarks, it is impossible to systematically assess whether systems can produce gender-neutral outputs, let alone determine when they do so appropriately. This Chapter addresses that gap by presenting evaluation resources developed to enable empirical research on gender-inclusive translation, responding to **RQ2**: *How can gender-inclusive translation be systematically represented and benchmarked?*

The creation of these resources was informed by a preliminary investigation of user perspectives on GNT (§4.1), which confirmed the acceptability of neutral forms and identified the communicative contexts where they are most appropriate. Building on these findings, §4.2 introduces GeNTE, the first natural benchmark for evaluating GNT from English into Italian. GeNTE is designed to test whether systems can neutralize when appropriate while preserving gender when explicitly marked in the source. The Chapter also presents its multilingual extension, mGeNTE, which expands coverage to German, Spanish, and Greek to enable cross-linguistic investigation. Finally, §4.3 introduces Neo-GATE, a benchmark for evaluating the use of neomorphemes in translation, addressing the evaluation needs of the innovative gender-inclusive strategies discussed in §3.1.3. Together, these resources provide the empirical foundation for the evaluation methods and generation experiments presented in subsequent Chapters.

4.1 User Perspectives on Gender-Neutral Translation

Before developing evaluation resources for GNT, a preliminary investigation was conducted to assess the acceptability of neutral forms among potential MT users. In fact, while institutional guidelines endorse gender-neutral language in formal communication (§3.1.1), the perception of such forms in cross-lingual settings remained underexplored. Prior work by Lardelli and Gromann [260] examined the cognitive effort that GNT requires of professional post-editors, but no study had investigated how a broader range of stakeholders perceive neutral translations compared to gendered alternatives. However, understanding user attitudes is essential for determining whether GNT represents a viable direction for MT development and for identifying the contexts in which neutral forms are most appropriate.

To address this gap, an online questionnaire was designed and distributed in April 2023. targeting individuals with high competence in both English and Italian. The survey was distributed via targeted emails and social media posts, with requests to share within relevant communities.

Participant selection and demographics. Participation was voluntary, uncompensated, and anonymous, with no identifying information collected. Participants were free to withdraw at any time without consequence.[1] Importantly, the survey did not target professional translators or MT specialists; rather, the goal was to gather perspectives from general stakeholders who might encounter MT outputs directly or indirectly, such as through automatically translated web content. Participants were informed that results would inform research on inclusive MT. Since the survey required judging English→Italian translations, only participants with high competence in both languages were eligible: Italian at C1 level or higher and English at B2 or higher, following the Common European Framework of Reference [103]. Screening questions verifying these language skills were placed at the beginning of the survey, along with an age verification excluding participants under 18. Of the 101 responses received, 98 were from eligible participants and thus included in the analysis.

The survey included a section collecting background information such as educational level, field of study, age, and self-reported gender identity. The participant pool was relatively homogeneous: the majority held a master's degree, and the most represented age range was 24–35. This homogeneity was expected given the distribution channels and the high English competence required for participation. Regarding gender, responses to the open question "How do you identify?" (Table 4.1) show a higher representation of women (57 responses, including 'woman,' 'female,' and "cisgender woman") compared to men (33 responses). This

[1] An archived version of the survey is accessible at: `https://forms.gle/YL76UeWbe4NWdCPPA`.

Response	Count
woman	55
man	29
I don't define myself	1
non-binary transgender	1
trans man	1
cisgender woman	1
cis male	1
female	1
male	2
lad	1
– (no response)	5

Table 4.1: Open responses to the question: *How do you identify?*

distribution likely reflects the voluntary nature of participation, which may have attracted individuals more engaged with the topic of inclusive language. Rather than viewing this as a limitation, this can be considered as an opportunity to gather perspectives from relevant stakeholders who are often most affected by discriminatory language practices [115].

Linguistic acceptability assessment. The core of the survey assessed linguistic acceptability through comparative judgments. Participants were presented with seven English source sentences paired with two Italian translation alternatives: a gendered translation (GT) using masculine generic forms and a neutral translation (NT) employing neutralization strategies. The original source sentences and their gendered translations were retrieved from EU multilingual documents in the administrative and legislative domain. A professional linguist with expertise in gender-inclusive language then created the neutral alternatives. This pairing methodology provided a fixed basis for comparison: presenting GT and NT side by side ensured that any preference differences could be attributed to gender-related factors rather than other aspects of translation quality that might influence acceptability judgments.

For each of the example sentences presented, participants indicated whether they preferred the GT, the NT, or found them equivalent. Follow-up questions probed the reasons behind their choices and gathered insights on perceived limitations of the neutral alternatives (Figure 4.1). Additionally, for three source sentences, participants selected their preferred neutral translation from four different options, providing data on preferences among different neutralization strategies.

Results on translation preferences. The results provide evidence on the relative acceptability of GNT. Aggregating across all examples, 42.5% of responses indicated a preference for the neutral translation, 36.5% found the two alternatives equivalent, and only 21% preferred

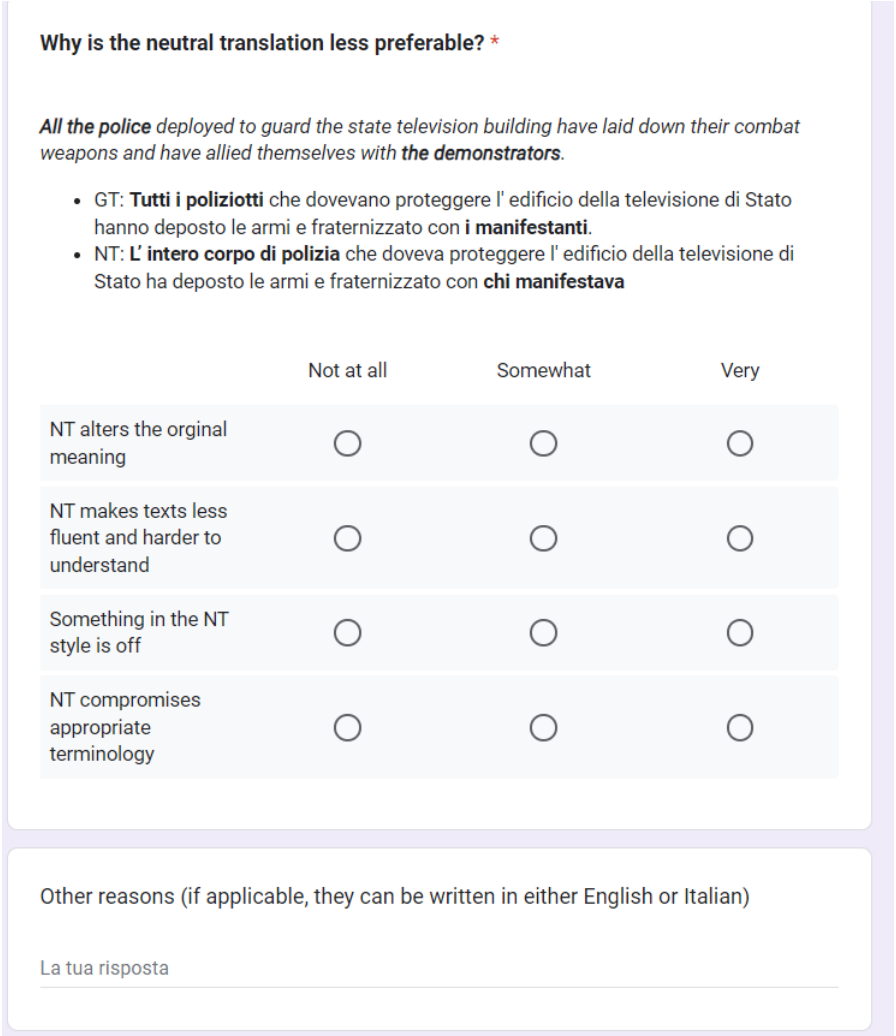

Figure 4.1: Questionnaire: follow-up questions on linguistic acceptability.

the gendered translation. This distribution suggests that neutral forms are not merely tolerated but are often actively preferred over masculine generics.

Disaggregating by example reveals variation across different neutralization strategies, consistent with the taxonomy presented in §3.1.2. Simple lexical substitutions, such as replacing *l'uomo* (man) with *gli esseri umani* (human beings), received particularly strong support (50% NT preference vs. 10.4% GT preference). More complex sentence reformulations were also considered acceptable, though responses varied depending on whether the reformulation affected perceived meaning or style. When neutral alternatives required longer periphrases or altered the sentence structure more substantially, participants showed greater ambivalence. However, even in such cases, outright preference for gendered translations remained a minority position.

Attitudes toward gender-neutral language. The final portion of the survey directly investigated participants' attitudes toward gender-neutral language use more broadly. To avoid influencing responses on the translation examples, these questions were placed after the linguistic acceptability section despite conceptually preceding it.

Responses to questions about use and acceptance revealed a clear distinction between formal and informal contexts. Participants reported both higher use and greater acceptance of neutral language in formal communication settings compared to informal ones, aligning with the institutional focus of the guidelines analyzed in §3.1.1. When asked about willingness to sacrifice different communicative aspects to ensure neutrality (Figure 4.2), participants showed notable flexibility: the majority indicated willingness to accept longer or stylistically different formulations if they achieved gender-neutrality, at least in formal contexts [480, 157]. This finding validates the viability of neutralization strategies that may introduce verbosity, as discussed in §3.1.2, particularly for formal and institutional domains [415].

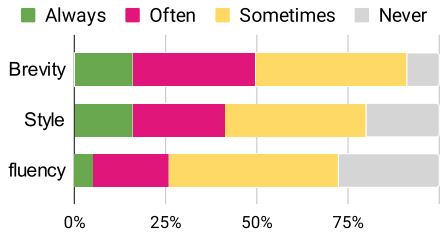

Figure 4.2: Willingness to sacrifice different communicative aspects to ensure neutrality.

Implications for benchmark design. These findings informed several design choices for the GeNTE corpus presented in the following Section. First, the overall acceptance of neutral translations validates GNT as a meaningful research direction worth systematic benchmarking. Second, the preference for neutral forms in formal contexts motivated the selection of Europarl [246] as the source corpus, as it represents precisely the institutional and administrative communication domain where neutral language is most accepted and expected. Third, the observation that different neutralization strategies elicit varying degrees of acceptance motivated the inclusion in GeNTE of multiple neutral references by independent translators, capturing the variability of valid solutions rather than prescribing a single correct neutralization for each sentence (see §4.2.2). Finally, the finding that acceptance varies by context reinforced the importance of including both *(i)* scenarios requiring neutralization, and *(ii)* scenarios where gender should be preserved, ensuring that benchmarks test appropriate application of neutral forms rather than indiscriminate neutralization.

Key Concepts

- **User Acceptability Study**: A preliminary investigation assessed the acceptability of neutral forms among potential MT users through an online questionnaire distributed to 98 participants with high competence in both English and Italian. Results indicate that neutral translations are not merely tolerated but often actively preferred: 42.5% of responses favored neutral translations, 36.5% found neutral and gendered alternatives equivalent, and only 21% preferred gendered forms.
- **Strategy and Context Variation**: Acceptance varies by neutralization strategy and communicative context: simple lexical substitutions receive particularly strong support, while complex reformulations show greater ambivalence. Participants report higher acceptance of neutral language in formal settings, with willingness to sacrifice brevity or style to achieve neutrality.
- **Implications for Benchmark Design**: These findings validate GNT as a research direction, motivate the selection of Europarl as a formal-domain source corpus, support the inclusion of multiple neutral references to capture strategy variability, and reinforce the need for dual-scenario evaluation (neutralization versus gender preservation).

4.2 The GeNTE Corpus

Building on the insights from the user survey, this work introduces GeNTE (**Ge**nder-**N**eutral **T**ranslation **E**valuation), the first benchmark designed to evaluate MT systems' ability to perform GNT.[2] The corpus consists of 1,500 English-Italian parallel sentences containing mentions of human referents, equally divided between two subsets:

- `Set-N` (750 sentences): featuring gender-ambiguous source sentences that should be rendered neutrally in translation (e.g., *We have become used to the practice, as the previous speaker just said.*);
- `Set-G` (750 sentences): featuring gender-unambiguous source sentences containing explicit gender cues, which should be properly rendered with gendered (masculine or feminine) forms in translation (e.g., *I would like the Minister to comment on this point, if she can.*).

[2]GeNTE is publicly available at `https://huggingface.co/datasets/FBK-MT/GeNTE`.

Together, these sets allow benchmarking of whether systems can perform GNT when appropriate and whether they overgeneralize the use of neutral forms [413]. The design principles underlying this dual-scenario structure are discussed in §4.2.1.

GeNTE is built on naturally occurring data from Europarl [246], selected for several reasons. First, Europarl is a widely used, high-quality MT resource that facilitates comparison with prior work [73, 62]. Second, it represents formal communicative situations from the administrative and institutional domain, precisely the context for which gender-neutral forms are traditionally intended and where the survey findings (§4.1) indicated the highest acceptance [200, 415]. Third, as examined by Saunders and Olsen [405], Europarl exhibits a large proportion of gender-ambiguous cases that, although translated with gendered forms in the original Italian references, are suitable candidates for neutralization. For each original Europarl sentence pair, an additional gender-neutral reference translation was created, enabling contrastive evaluation of system outputs against both gendered and neutral alternatives.

4.2.1 Corpus Design Principles

The dual-scenario design of GeNTE operationalizes the desiderata for GNT articulated in §3.2. The first desideratum (D1) establishes that **neutralization should apply when gender cannot be reliably inferred from the source**. `Set-N` implements this principle by including sentences where English source texts lack explicit gender markers for human referents. These comprise gender-ambiguous terms such as *citizens*, *teachers*, and *colleague*, as well as masculine generics like *man* and compounds such as *chairman* and *spokesman* [303]. Following desideratum D3, masculine generics are treated as unreliable gender cues whose propagation to the target language would conflict with inclusive MT goals. Accordingly, they are included in `Set-N` as candidates for neutralization rather than in `Set-G` as sources of gender information. The second desideratum (D2) specifies that **neutralization should not override explicit gender information present in the source**. `Set-G` implements this by including sentences where the source contains unambiguous gender cues: lexically gendered words (*sister*, *woman*, *brother*), gendered titles (*Mr*, *Mrs*, *Ms*), and marked pronouns (*him*, *her*, *he*, *she*). For these sentences, appropriate translation requires preserving the gender expressed in the source rather than neutralizing it.

This dual-scenario design distinguishes GeNTE from prior gender bias benchmarks. Resources such as WinoMT [441], and MT-GenEval [107] focus exclusively on binary gender accuracy, testing whether systems produce the correct masculine or feminine form when gender is specified. GeNTE extends this paradigm by also testing whether systems can recognize when gender should *not* be specified, producing neutral outputs instead. This enables a more

comprehensive assessment aligned with the goals of gender-inclusive translation.

4.2.2 Data Collection and Creation

The construction of GeNTE involves three phases: automatic extraction of candidate sentences from Europarl, manual editing to ensure corpus consistency and annotation, and professional creation of neutral reference translations.

Data extraction. Europarl segments representing both translation scenarios are retrieved through regular expressions targeting specific linguistic patterns.[3] The extraction criteria were designed to: *(i)* identify source sentences containing mentions of human referents, *(ii)* maximize the variability of linguistic phenomena included in the corpus, and *(iii)* ensure balanced representation of both ambiguous and unambiguous gender cases.

For `Set-G`, the extraction targeted source sentences containing explicit gender cues: lexically gendered nouns (*sister*, *woman*, *brother*), titles (*Mr*, *Mrs*, *Ms*), and marked pronouns (*him*, *her*, *he*, *she*). For `Set-N`, the extraction matched word classes that do not convey gender distinctions in English but typically correspond to gendered expressions in Italian. These include second-person pronouns (*you*), occupational terms (*citizens*, *teachers*, *workers*, etc.). The extraction also targeted masculine generics such as *man* and its compounds (*chairman*, *layman*, *spokesman*), treating these as candidates for neutralization rather than as reliable masculine gender cues.

Sentence editing. The automatically extracted sentences required manual editing to ensure consistency and facilitate evaluation. Some source sentences contained mentions of multiple referents requiring different gender treatments in translation (e.g., a gender-ambiguous group and a specifically gendered individual). Such sentences were edited to include only referents requiring the same type of form, either all neutral or all gendered (with the same gender). This intervention ensures that each sentence pair can be evaluated as a coherent unit, prioritizing systematic evaluation over breadth of coverage. In fact, real-world source sentences can contain multiple human referents requiring different gender expression in the target, which in turn requires annotation and generation decisions at the word or entity level rather than the sentence level. The controlled, homogeneous constructions in GeNTE thus represent a streamlined scenario that facilitates systematic evaluation while leaving more complex, multi-referent configurations as a direction for future work (§8.3.4).

[3]Europarl data was retrieved from `https://www.statmt.org/europarl/archives.html`.

A second intervention addressed the under-representation of unambiguous feminine cases in the original Europarl data, a finding that confirms prior observations about gender imbalance in parliamentary corpora [405, 112], also due to the historically imbalanced representation of gender groups in that context [475]. To achieve balanced representation in `Set-G`, gender-swapping was performed on a subset of masculine instances, converting masculine markers to feminine equivalents [535, 53]. While these edits slightly reduce the naturalness of some examples, they enable sound evaluation by ensuring that performance differences between masculine and feminine cases can be meaningfully compared.

Additional minor edits addressed quality issues such as typos, translation errors in the original Europarl references, and overly complex sentences that would complicate evaluation. All editing interventions are documented in Appendix B.1.

Annotation. Following the editing phase, all sentence pairs were annotated as N (neutral) for `Set-N`, or as M (masculine) or F (feminine) for `Set-G`. This annotation process also verified that automatically extracted candidates were correctly assigned to their respective sets by examining the sentence context. For instance, gendered pronouns used as masculine generics (e.g., *It is up to an accused employer to prove **his** innocence*) were identified as N, while the same pronouns used referentially (e.g., *I would like to thank Commissioner Byrne for **his** cooperation*) were identified as M. While entity-level annotation schemes, such as gENder-IT [477], offer a finer-grained distinction between known and unknown gender per referent, sentence-level annotation is more appropriate for GeNTE given the diversity of neutralization strategies involved. In particular, some strategies (e.g., omission and impersonal formulations, respectively E and H in Table 3.2) remove referent mentions altogether (e.g., IT *tutti*$_{\text{[M]}}$ *dobbiamo agire*$_{\text{[we all must act]}}$ rendered as *si deve agire*$_{\text{[it is necessary to act]}}$), making consistent and aligned entity-level tagging non-trivial.

Creation of Neutral References. Confirming the predominant use of gendered forms when translating into grammatical gender languages, 97.2% of the segments collected from Europarl contained gendered Italian references. To implement the dual-scenario evaluation design outlined in §4.2.1, an additional gender-neutral reference was created for each sentence pair enabling contrastive assessment of system outputs against both gendered and neutral alternatives. This design follows the approach of binary gender bias benchmarks such as MuST-SHE [53], MT-GenEval [107], and GATE [366], which pair masculine and feminine references to isolate gender as the variable of interest. In this case, pairing gendered and neutral references allows isolation of gender-related linguistic elements as the source of variation when evaluating system outputs.

Recognizing that neutralization is an open-ended task with high variability in valid solutions (§3.1.2), three professional translators were engaged through a translation agency to create the neutral references.[4] Each translator was assigned a distinct portion of the Italian references to post-edit, replacing gendered terms with neutral formulations while preserving meaning and fluency. An expert linguist, native speaker of Italian and experienced with gender-inclusive language, prepared detailed neutralization guidelines drawing from institutional resources for the administrative domain.[5] After an initial training session, the linguist provided ongoing support throughout the process and reviewed all neutralizations for quality assurance. Qualitative insights from this revision process are provided in Appendix B.

It is worth noting that while `Set-G` sentences are unambiguously gendered, neutral reference translations are also constructed for these sentences to support the contrastive evaluation protocol described in §5.1. This means that, by design, neutralization is attempted even for lexically gendered terms such as *sister* or *daughter*, where gender is inherent to the word's meaning. These cases are included not because neutralization is linguistically expected or natural, but because the contrastive setup requires a neutral counterpart for every gendered reference.

4.2.3 The COMMON-SET and Linguistic Variability

While each translator was responsible for neutralizing a distinct portion of the corpus, a subset of 200 sentences was also designated to be neutralized by all three translators independently. This COMMON-SET, containing 100 sentences from `Set-N` and 100 from `Set-G`, yields 200 source sentences each paired with one gendered reference and three neutral references. The COMMON-SET serves two purposes: it provides a subset for testing the robustness of evaluation protocols across different neutral references (§5.1), and it allows measurement of linguistic variability among neutralization solutions.

Table 4.2 illustrates this variability with examples from the COMMON-SET. Example *i* demonstrates the range of solutions translators produced for the same gendered expression: *tutti i miei colleghi* (all my colleagues$_{[M]}$) was neutralized as *agli altri membri* (to other members), *ogni collega* (each colleague), and *tutte le persone con cui lavoro* (all the people with whom I work). These solutions employ different strategies from the taxonomy in §3.1.2: plural and singular epicenes and periphrasis respectively. Example *ii* shows a rarer case where all translators converged on the same solution, replacing the gendered *del collega* (of the colleague$_{[M]}$) with *dell'onorevole collega* (of the honorable colleague). Such convergence

[4]Translators were compensated at 60 euros per hour, with each translator working approximately 14 hours.
[5]The guidelines are released together with the GeNTE corpus.

occurred when a single neutralization strategy was clearly optimal for the context. Example *iii* illustrates cases that posed particular challenges. The gendered terms *sorella* (sister) and *Commissaria* (Commissioner$_{[F]}$) require verbose periphrases to neutralize, compromising the original text's fluency and style. Two of the three translators judged neutralization infeasible for this sentence and did not provide a neutral reference.

i – N	**SRC**	I, along with **all my colleagues**, wish to welcome this [...]
	REF-G	Insieme a **tutti i miei colleghi**, desidero esprimere il mio compiacimento per questa [...]
	REF-N 1	Insieme **agli altri membri**$_{\texttt{[other members]}}$, desidero esprimere il mio compiacimento per questa [...]
	REF-N 2	Insieme a **ogni collega**$_{\texttt{[each colleague]}}$, desidero esprimere il mio compiacimento per questa [...]
	REF-N 3	Insieme a **tutte le persone con cui lavoro**$_{\texttt{[all the persons with whom I work]}}$, desidero esprimere il mio compiacimento per questa [...]
ii – M	**SRC**	I welcome this excellent report **from my colleague** Mr Skinner.
	REF-G	Valuto positivamente la relazione **del collega**, onorevole Skinner.
	REF-N 1	Valuto positivamente la relazione **dell'onorevole collega**$_{\texttt{[of the honorable colleague]}}$ Skinner.
	REF-N 2	Valuto positivamente la relazione **dell'onorevole collega** Skinner.
	REF-N 3	Valuto positivamente la relazione **dell'onorevole collega** Skinner.
iii – F	**SRC**	Mrs Ana de Palacio Vallelersundi has **a sister who is a Commissioner** [...]
	REF-G	**La** onorevole [...] ha **una sorella, la quale è una Comissaria** [...]
	REF-N 1	N.A.
	REF-N 2	L'onorevole [...] ha **uno stretto legame di parentela**$_{\texttt{[is closely related]}}$ con **un membro della Commissione**$_{\texttt{[a member of the Commission]}}$
	REF-N 3	N.A.

Table 4.2: Examples of entries in the COMMON-SET. REF-G indicates the gendered references, REF-N 1, 2, 3 indicate the neutralized references produced by Translator 1, 2, and 3 respectively. Words in **bold** are mentions of human referents; underlined words are linguistic cues informing about the referents' gender.

Analysis of the COMMON-SET quantifies this variability. The three translators produced identical neutral references in only 13.57% of cases, with an additional 8% showing high similarity (e.g., the same neutral words in different order). The remaining approximately 79% of cases exhibit substantial variability in neutralization solutions. Appendix B.3 reports further analyses of the variability within the gender-neutral references in `Set-N`.

These statistics carry important implications for both evaluation methodology and generation research. The high variability empirically confirms the open-ended nature of neutralization discussed in §3.1.2: the same gendered expression can be legitimately neutralized by means of different strategies, each yielding different surface forms.

This variability poses a fundamental challenge for reference-based evaluation. Standard MT metrics that reward similarity to a single reference will penalize valid neutralizations that happen to differ from that reference [69, 373], potentially misinterpreting successful GNT as low-quality translation [147, 296, 21]. The COMMON-SET, with its 600 neutral references for 200 source sentences, provides a controlled subset for testing evaluation protocols' robustness across different valid neutralizations, while its balanced distribution between `Set-N` and `Set-G`, and between masculine and feminine instances within `Set-G`, supports unbiased assessment of system performance across all scenarios.

At the same time, the approximately 21% of cases where translators independently converged on identical or highly similar solutions suggests that neutralization is not completely unpredictable. Certain contexts appear to favor particular strategies: when a single neutralization is clearly optimal for the register, meaning preservation, and fluency constraints, trained translators recognize and select it. This convergence indicates that neutralization follows learnable patterns grounded in linguistic principles, offering encouragement that MT systems might acquire similar capabilities given appropriate training signals or prompting strategies.

4.2.4 Corpus Statistics and Characteristics

Table 4.3 presents the statistics for GeNTE and its COMMON-SET. The gendered references (REF-G) contain 4,263 gendered words requiring neutralization: 1,972 in `Set-N` and 2,148 in `Set-G` (after gender-balancing interventions). The high count of gendered words in `Set-N` confirms that the original Europarl translations extensively employ masculine generics for gender-ambiguous referents, validating the corpus as a suitable testbed for GNT evaluation.

The neutral references (REF-N) are slightly longer on average than their gendered counterparts (26.95 vs. 24.66 words in `Set-N`; 26.55 vs. 25.26 in `Set-G`), reflecting the verbosity that some neutralization strategies introduce. This length difference, while modest, is consistent with the survey finding that users are willing to accept slightly longer formulations to achieve neutrality (§4.1).

4.2.5 The mGeNTE Multilingual Extension

GeNTE established the first benchmark for GNT evaluation in English→Italian, providing a foundation for systematic research on gender-inclusive translation. However, the challenges of GNT are not uniform across languages: different grammatical gender systems offer distinct neutralization resources, and strategies effective in one language may be unavailable or inappropriate in another. Understanding these cross-linguistic variations is essential for

	Source		REF-G			REF-N	
	Entries	Avg length	Entries	Avg length	# Gendered words	Entries	Avg length
				GENTE			
Set-N	750	25.67	750	24.66	1,972	750	26.95
Set-G	750	26.51	750	25.26	2,148	750	26.55
				COMMON-SET			
Set-N	100	27.45	100	27.00	300	300	28.87
Set-G	100	26.99	100	26.34	299	300	27.57

Table 4.3: Corpus statistics for GENTE and its COMMON-SET. Both sets requiring gendered translations (Set-G) are equally balanced between feminine and masculine sentences. Average lengths are calculated excluding punctuation. The *Gendered words* column reports the total number of words in the REF-Gs that required neutralization in the REF-Ns.

developing MT systems that can produce inclusive outputs regardless of the target language. To enable such investigation, this Section presents mGeNTE, an extension of the original benchmark to cover English→German, English→Spanish, and English→Greek in addition to Italian.[6]

The target languages were selected to represent diverse grammatical gender systems [97, 439, 188]. Italian and Spanish, as Romance languages, share extensive gendered morphology on nouns, adjectives, articles, and some verbal forms [193, 293]. German, from the Germanic branch, presents a three-gender system where the neuter gender sometimes offers neutralization possibilities unavailable in Romance languages, though human referents are still typically marked as masculine or feminine [200]. Greek was deliberately included as a lower-resource language with a distinct script [205], thereby broadening linguistic diversity and ensuring coverage of underrepresented cases in inclusive MT research. Together, these languages enable systematic comparison of how GNT challenges vary across typologically related but morphologically distinct systems, as illustrated in Table 4.4.

mGeNTE preserves the original GeNTE design rationale and curation methodology to ensure comparability across language pairs. Each language pair comprises 1,500 sentence triplets consisting of an English source, a gendered target reference from Europarl, and a newly created gender-neutral target reference, resulting in 6,000 total entries across the four languages. The corpus maintains the dual-scenario structure established for GeNTE (§4.2.1), with 750 sentences per language in each subset (Set-G and Set-N), enabling evaluation of appropriate GNT application across a diverse set of target languages.

[6] mGeNTE is publicly available at https://huggingface.co/datasets/FBK-MT/mGeNTE.

Set-N	SRC	**Pensioners** are in favour of strengthening criminal law, [...]
en-it	REF-G	**I pensionati** sono favorevoli a un rafforzamento del diritto penale, [...]
	REF-N	**Le persone pensionate**$_{\texttt{[pensioned people]}}$ sono favorevoli a un rafforzamento del diritto penale, [...]
en-es	REF-G	**Los pensionistas** están a favor de reforzar el Derecho penal no solo nacional, [...]
	REF-N	**Hay pensionistas**$_{\texttt{[there are pensioners]}}$ que están a favor de reforzar el Derecho penal no solo nacional, [...]
en-de	REF-G	Die **Rentner** begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]
	REF-N	**Die Menschen in Rente**$_{\texttt{[people in retirement]}}$ begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]
en-el	REF-G	Οι συνταξιούχοι είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]
	REF-N	Τα συνταξιοδοτημένα άτομα$_{\texttt{[the retired individuals]}}$ είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]
Set-G	SRC	I trust the **Commissioner** will promise that <u>he</u> will exercise extra vigilance.
en-it	REF-G	Spero che **il Commissario** ora prometta di vigilare attentamente a tale riguardo.
	REF-N	Spero che **il membro della Commissione**$_{\texttt{[the member of the board]}}$ ora prometta di vigilare attentamente a tale riguardo.
en-es	REF-G	Espero que **el Comisario** prometa controlar exhaustivamente esta situación.
	REF-N	Espero que **la representación de la Comisión**$_{\texttt{[the representative of the board]}}$ prometa controlar exhaustivamente esta situación.
en-de	REF-G	Von **dem Herrn Kommissar** erwarte ich heute die Zusage, **er** werde mit Argusaugen darüber wachen.
	REF-N	**Von dem Kommissionsmitglied**$_{\texttt{[From the board member]}}$ erwarte ich heute die Zusage, **es**$_{\texttt{[it]}}$ werde mit Argusaugen darüber wachen.
en-el	REF-G	Προσδοκώ από τον Επίτροπο να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.
	REF-N	Προσδοκώ από το μέλος της Επιτροπής$_{\texttt{[the member of the Commission]}}$ να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.

Table 4.4: Examples from the mGeNTE PARALLEL-SET, showing entries from `Set-N` and `Set-G` with their gendered (REF-G) and neutral (REF-N) references across all four language pairs. Words in **bold** are mentions of human referents; <u>underlined</u> source words are explicit gender cues.

Multilingual Data Extraction. The construction of mGeNTE prioritized alignment with the original GeNTE corpus to enable direct cross-linguistic comparison. It began by retrieving Europarl sentences corresponding to the original GeNTE entries using their Europarl identifiers, extracting parallel versions in German, Spanish, and Greek. Each automatically extracted sentence was manually verified to confirm that it contained a gender-related phenomenon in

the target language; sentences where gender marking did not apply were discarded.[7] This process yielded 987 sentences that are fully parallel across all four language pairs.

To reach the target of 1,500 sentences per language, the remaining entries were extracted independently for each language using regular expressions targeting language-specific gender phenomena. For `Set-G`, the extraction matched unambiguous English sources containing explicit gender cues (titles, pronouns, gendered nouns). For `Set-N`, the extraction excluded segments with source gender information and matched expressions that, while gender-ambiguous in English, correspond to gendered forms in the respective target language. The specific patterns vary by language: for instance, *deputy* maps to gendered forms in Spanish (*diputado*$_{[M]}$/*diputada*$_{[F]}$), German (*Stellvertreter*$_{[M]}$/*Stellvertreterin*$_{[F]}$), and Greek (αναπληρωτής$_{[M]}$/αναπληρώτρια$_{[F]}$).

Sentence Editing and Balancing. Following the GeNTE methodology, the extracted sentences underwent manual editing to ensure corpus consistency. Sentences containing multiple referents requiring different gender treatments were edited to include only referents requiring the same form type, allowing each entry to be evaluated as a coherent unit. To address the under-representation of feminine forms in the original Europarl data, gender-swapping interventions achieved balanced masculine-feminine distribution within `Set-G`. The number of editing interventions varied by language: 652 for German, 621 for Spanish, and 702 for Greek, with approximately 60% serving gender balancing and the remainder addressing quality issues. Minor corrections for typos and translation errors were also applied. Following the editing process, all sentences were manually classified as N for `Set-N`, or as M (masculine) or F (feminine) for `Set-G`.

Creation of Neutral References. The creation of gender-neutral references was entrusted to professional translators with native or C1-level competence in their assigned target language and specialized training in GNT. Translators post-edited the gendered Europarl references, replacing gendered terms with neutral formulations following language-specific guidelines adapted from the original GeNTE instructions. These guidelines drew from institutional resources for inclusive language in each target language, accounting for the different neutralization strategies available across grammatical gender systems. The resulting neutral references capture the variability of valid solutions (see Table 4.4) while maintaining consistency with the neutralization principles established for the original Italian corpus.

[7] Cross-linguistic differences in gender encoding can result in a sentence being relevant for one language pair but not another. For instance, English *child* corresponds to gender-marked forms in Italian (*il/la bambino/a*) but remains invariant in German (*das Kind*).

The PARALLEL-SET. Among the 1,500 entries per language pair, 987 source sentences are fully aligned across all four languages, with an additional subset of 578 sentences (the PARALLEL-SET) verified as containing comparable gender phenomena across all pairs. This parallel subset enables direct cross-linguistic analysis: the same English source can be evaluated against Italian, Spanish, German, and Greek outputs, revealing how model behavior varies across target languages for identical inputs. Table 4.4 illustrates entries from the PARALLEL-SET, showing how the same English source is rendered across all four language pairs, whereas Table 4.5 summarizes the distribution of mGeNTE segments across language pairs and subsets.

	Set-G	Set-N	Gendered words	Unique words
EN→IT	750	750	4,115	802
EN→ES	750	750	4,363	644
EN→DE	750	750	3,977	613
EN→EL	750	750	3,736	743
PARALLEL-SET	409	578	–	–

Table 4.5: Distribution of mGeNTE segments by subset and language pair, including sentences fully parallel across all pairs (PARALLEL-SET). The rightmost columns report total and unique annotated gendered words per language.

Gendered Word Annotations. To enable fine-grained evaluation and cross-linguistic comparison, all target sentences in mGeNTE were manually annotated at the word level to identify gendered terms requiring neutralization. This annotation layer, absent from the original GeNTE release, was added retrospectively to the Italian data and created alongside the new language pairs. The annotations identify all words expressing grammatical gender in reference to human beings, including nouns, adjectives, articles, pronouns, and verbal forms where applicable. The bold words in Table 4.4 exemplify such annotations.

To ensure annotation reliability, a second annotator independently reviewed a subset of target sentences for each language. Inter-annotator agreement, measured using the Dice coefficient [120, 496] for overlap between annotated word sets, exceeded 0.92 across all languages, indicating high consistency. Disagreements were primarily oversights rather than substantive differences in annotation criteria and were reconciled through discussion.

Table 4.5 reports the total and unique counts of annotated gendered words per language. Spanish exhibits the highest total count (4,363), followed by Italian (4,115), German (3,977), and Greek (3,736). The differences reflect both the morphological characteristics of each language and the specific sentences included in each subset. Notably, Italian and Greek show higher counts of unique gendered words (802 and 743 respectively) compared to Spanish

and German (644 and 613). This pattern reflects morphosyntactic differences: Italian marks gender on verbal participles, while Greek employs gendered plural pronouns, both contributing additional unique forms not present in the other languages.

Cross-Linguistic Observations. Qualitative analysis of the gendered word annotations reveals patterns relevant to GNT research. The `Set-N` annotations are heavily populated with masculine forms used generically to refer to mixed or unknown referents. For instance, Italian *i cittadini* (the citizens$_{[M]}$), Spanish *los trabajadores* (the workers$_{[M]}$), German *die Abgeordneten* (the deputies$_{[M]}$), and Greek οι πολίτες (the citizens$_{[M]}$) all appear frequently in contexts where the referents' gender is unspecified. These cases, translated with masculine forms in the original Europarl references, represent precisely the target of language neutrality efforts [186] and validate mGeNTE's suitability for GNT evaluation.

Table 4.4 also illustrates the range of neutralization strategies applied across languages for identical source sentences. The `Set-N` example shows how the gender-ambiguous English term *pensioners* is neutralized differently in each language: Italian employs a participial construction (*le persone pensionate*), German uses a prepositional phrase (*die Menschen in Rente*), Greek shifts to a neuter noun phrase (τα συνταξιοδοτημένα άτομα), while Spanish restructures the sentence to avoid the gendered subject position (*hay pensionistas que...*). These solutions reflect the different linguistic resources available in each language for achieving neutrality.

The annotations also reveal cross-linguistic differences in how gender permeates the grammatical system. Spanish stands out for its gendered plural pronouns across all persons: first person (*nosotros*$_{[M]}$/*nosotras*$_{[F]}$), second person (*vosotros/vosotras*), and third person (*ellos*$_{[M]}$/*ellas*$_{[F]}$) all distinguish masculine from feminine, requiring neutralization even for basic pronominal reference. Greek similarly marks gender in the third-person plural (αυτοί/αυτές/αυτά for masculine/feminine/neuter), though first and second person plurals remain invariant. In contrast, Italian and German do not mark gender on plural pronouns: Italian *noi*, *voi*, and the common *loro*[8] are gender-invariant, as are German *wir*, *ihr*, and *sie*.

The `Set-G` example in Table 4.4 further illustrates how explicit gender marking in the source interacts with translation across languages. While Italian and Spanish neutralize the gendered noun (*Commissario* → *membro della Commissione*; *Comisario* → *representación de la Comisión*), the pronoun coreference is resolved differently in German: the neuter pronoun *es* refers to *Kommissionsmitglied*, exploiting the grammatical gender of the neutral noun rather than preserving the source's masculine reference.

[8]Technically, Italian does feature gendered pronouns for the third person plural: *essi*$_{[M]}$ and *esse*$_{[F]}$. However, these are considered obsolete in contemporary language.

These differences indicate that a sentence requiring neutralization of plural pronouns in Spanish or Greek may pose no equivalent challenge in Italian or German. More broadly, a system that successfully performs GNT for one language pair may struggle with another due to differences in available neutralization strategies, morphological complexity, or the distribution of gendered forms. The mGeNTE benchmark enables systematic investigation of such variations, supporting the development of inclusive MT systems that generalize across typologically diverse languages. The evaluation methods leveraging mGeNTE for cross-linguistic GNT assessment are presented in §5.3.

Key Points

- **The GeNTE Corpus**: The first evaluation resource for GNT, comprising 1,500 English-Italian sentence pairs from Europarl with dual references (gendered and neutral). Its dual-scenario design operationalizes the desiderata from §3.2: `Set-N` tests neutralization of gender-ambiguous sources, while `Set-G` tests preservation of explicit gender cues.
- **Corpus Construction**: GeNTE was built through three phases: automatic extraction of candidate sentences from Europarl using regular expressions targeting gender-related patterns, manual editing to ensure consistency (including gender-swapping to balance masculine and feminine representation), and professional creation of neutral references by three translators following dedicated guidelines.
- **mGeNTE Multilingual Extension**: An extension covering English to Italian, Spanish, German, and Greek (6,000 total entries), enabling cross-linguistic investigation of GNT. Word-level annotations of gendered terms support fine-grained evaluation, while the PARALLEL-SET (578 sentences identical across all language pairs) enables direct comparison of system behavior across typologically diverse target languages.
- **Cross-Linguistic Variation**: Different languages offer distinct neutralization resources and challenges. Spanish and Greek mark gender on plural pronouns, whereas Italian and German do not; German's neuter gender sometimes provides neutralization options unavailable in Romance languages. These differences indicate that GNT strategies effective for one language may not transfer directly to another.

4.3 The Neo-GATE Benchmark

The benchmarks presented thus far address conservative gender-inclusive strategies, evaluating whether systems can produce translations that avoid gendered forms through established linguistic resources. However, as discussed in §3.1.3, gender-inclusive language also encompasses innovative approaches used to convey neutrality or explicitly represent non-binary identities through neologistic devices. Among these, neomorphemes have gained particular traction in Romance languages: novel characters or character combinations that replace gendered morphemes to create forms that are neither masculine nor feminine (e.g., IT *scienziatə* instead of *scienziato*$_{[M]}$ or *scienziata*$_{[F]}$; EN *scientist*).

To enable evaluation of MT systems' ability to generate translations with neomorphemes as the ones we discuss in §6.3, we developed Neo-GATE, a benchmark specifically designed for neomorpheme-based gender-inclusive translation from English into Italian.[9] Following the open-class conceptualization proposed by Lauscher et al. [261] for neopronouns, neomorphemes are treated as an evolving set of forms rather than a fixed inventory. This perspective is essential given the developing nature of neomorpheme usage: multiple paradigms co-exist (e.g., the schwa *ə/ɜ*, the asterisk *, the at-sign @, etc.), and communities continue to develop new forms [457, 447, 1]. Accordingly, Neo-GATE is designed to be adaptable to any neomorpheme paradigm through a flexible annotation system based on placeholder tags.

4.3.1 From GATE to Neo-GATE

Neo-GATE builds upon GATE [366], an existing benchmark for evaluating gender bias in MT. GATE provides English source sentences paired with Italian references that differ only in the gender of human referents: each entry includes both a masculine reference (REF-M) and a feminine reference (REF-F). Source sentences in GATE are designed to be gender-ambiguous: they contain no linguistic elements providing gender information about human referents, making gender assignment in translation entirely dependent on system behavior rather than source constraints or information. This design makes GATE an ideal foundation for our benchmark. Since neomorphemes are intended for human referents whose gender is unknown or who do not identify within the binary, GATE's gender-ambiguous sources represent precisely the contexts where neomorpheme usage would be warranted. Neo-GATE extends GATE by adding a third reference featuring neomorphemes, along with word-level annotations that identify all gender-marked terms in the references requiring transformation.

Neo-GATE includes all entries from GATE's test set, with two exceptions where the

[9] Neo-GATE is publicly available at `https://huggingface.co/datasets/FBK-MT/Neo-GATE`.

references contained no gender-marked terms relevant to human referents.[10] The resulting benchmark contains 841 test entries. Additionally, we annotated 100 entries from GATE's development set to provide exemplar sentences for few-shot prompting experiments (see §6.3).

Annotation Process. For each GATE entry, a tagged reference is created by systematically identifying all words expressing grammatical gender in reference to human beings and replacing their gendered components with placeholder tags. A linguist with expertise in Italian morphology performed the initial annotation following dedicated guidelines, which specified criteria for identifying gendered terms and rules for tag assignment.[11] A second linguist independently annotated a randomly selected 15% subset of the test set using the same guidelines. Inter-annotator agreement, computed with Cohen's kappa [92] on placeholder tag assignment, reached 0.94, indicating almost perfect agreement [257]. The few disagreements were attributable to oversights rather than substantive differences in interpretation and were reconciled.

4.3.2 Annotation Scheme and Paradigm Flexibility

The annotation scheme was designed to cover all parts of Italian grammar that express gender in reference to human beings while remaining adaptable to different neomorpheme paradigms. The scheme employs placeholder tags that can be automatically replaced with the appropriate forms for any desired paradigm before evaluation.

Tagset Design and Adaptation. The tagset distinguishes between content words and function words, and between singular and plural forms. For content words (nouns, adjectives, past participles), only the inflectional morpheme is replaced with a tag: `<ENDS>` for singular and `<ENDP>` for plural. For example, *direttore*$_{[M]}$/*direttrice*$_{[F]}$ (EN: director) becomes *direttor*`<ENDS>`.

Function words (articles, prepositions, pronouns) require whole-word replacement rather than morpheme substitution. This design choice reflects two characteristics of Italian grammar. First, some function words are not morphologically derived but paradigmatically opposed: the masculine singular definite articles *il* and *lo* versus the feminine *la* do not share a common stem. Second, competing neomorpheme forms exist for the same function word and may differ in the root: for instance, both *lɜ* and *ə* have been proposed for the plural definite article.[12]

[10] These entries were excluded as they provided no evaluation signal for neomorpheme generation.

[11] The annotation guidelines are released with Neo-GATE.

[12] The first form appears in `https://italianoinclusivo.it/scrittura/`; the second one in `https://effequ.it/schwa/`.

GATE	**Source**	The department chair said they might hire new professors
	Ref. Masc.	Il direttore del dipartimento ha detto che potrebbero assumere nuovi professori
	Ref. Fem.	La direttrice del dipartimento ha detto che potrebbero assumere nuove professoresse
NEO-GATE	**Ref. tagged**	<DARTS> direttor<ENDS> del dipartimento ha detto che potrebbero assumere nuov<ENDP> professor<ENDP>
	Annotation	il la <DARTS> dirett=1; direttore direttrice direttor<ENDS>; nuovi nuove nuov<ENDP> professor=1; professori professoresse professor<ENDP>
NEO-GATE *	**Reference**	L* direttor* del dipartimento ha detto che potrebbero assumere nuov* professor*
	Annotation	il la l* dirett=1; direttore direttrice direttor*; nuovi nuove nuov* professor=1; professori professoresse professor*
NEO-GATE ə/ɜ	**Reference**	Lə direttorə del dipartimento ha detto che potrebbero assumere nuovɜ professorɜ
	Annotation	il la lədirett=1; direttore direttrice direttorə; nuovi nuove nuovɜ professor=1; professori professoresse professorɜ

Table 4.6: Example of a single entry in GATE, NEO-GATE with placeholder tags, and NEO-GATE adapted to two neomorpheme paradigms. Terms relevant for evaluation are highlighted.

Whole-word placeholders accommodate this variation without prescribing a particular solution.

The complete tagset comprises 29 tags covering definite and indefinite articles (<DARTS>, <DARTP>, <IART>), partitive articles (<PARTP>), articulated prepositions with various roots (<PREPdiS>, <PREPaP>, etc.), demonstrative adjectives (<DADJquelS>, <DADJquestP>), possessive adjectives for all persons (<POSS1S> through <POSS4P>), and direct object pronouns (<PRONDOBJS>, <PRONDOBJP>). The full tagset with mappings to masculine, feminine, and neomorpheme forms is provided in Appendix C.1.

Before evaluation, the placeholder tags are replaced with forms from the desired neomorpheme paradigm. Table 4.6 illustrates this process for two paradigms used in the generation experiments discussed in §6.1: the asterisk paradigm, which uses * for both singular and plural, and the schwa paradigm, which distinguishes singular (*ə*) from plural (*ɜ*). The same tagged reference can be adapted to either paradigm, or to any future paradigm that may emerge, by defining the appropriate mappings.

Function Word Anchoring. To enable precise evaluation of function words, the annotation includes anchoring information that links each function word to its associated content word. This is necessary because Italian sentences may contain multiple function words of the same type (e.g., multiple definite articles), and evaluation must verify that each function word agrees with the correct noun rather than with other nouns appearing elsewhere in the sentence.

An anchor is defined as the longest common substring shared by the masculine, feminine, and tagged forms of the content word to which the function word is syntactically linked. The annotation format appends the anchor and a distance value to the function word forms: for

instance, *lo la l* student=1* indicates that the article forms (*lo*, *la*, or *l**) should be evaluated only when the substring *student* appears immediately after them (distance of 1 word). This ensures that the evaluation matches each article to its corresponding noun.

The distance parameter accommodates cases where other words intervene between the function word and its associated content word. For example, in a sentence like "let your friends see you" (REF-M: *lascia che i tuoi amici ti vedano*; REF-F: *lascia che le tue amiche ti vedano*; reference adapted to the '*' neomorpheme: *lascia che l* tu* amic* ti vedano*), the annotation *i le l* amic=2* specifies that the article should be found two words before the anchor *amic-*, accounting for the possessive adjective. The possessive itself is annotated separately as *tuoi tue tu* amic=1*, linking it to the same anchor at distance 1. This mechanism enables fine-grained evaluation even in syntactically complex constructions where multiple function words modify the same noun. Crucially, the anchoring system allows each annotated word to be evaluated independently: rather than requiring a complete matching sequence (e.g., *l* tu* amic**), the evaluation can recognize partial successes where only some words in a construction feature neomorphemes (e.g., *i tu* amici*, where only the possessive is correctly realized). This granularity provides a more nuanced assessment of model behavior, distinguishing between complete failures and cases where models demonstrate partial competence with neomorpheme generation.

Corpus Statistics. Table 4.7 presents statistics for Neo-GATE's test and development sets. The test set contains 841 entries with 2,479 annotated tags, of which 1,539 are content word tags and 940 are function word tags. The distribution between singular (1,316) and plural (1,163) forms is relatively balanced, ensuring that evaluation covers both morphological contexts. The development set of 100 annotated entries provides sufficient material for constructing few-shot prompts (as done in §6.3) while reserving the majority of data for evaluation. The evaluation metrics specifically designed for assessing neomorpheme generation are presented in §5.4, where they complement the general GNT evaluation methodologies.

	Entries	**Tags**	**Content**	**Function**	**Singular**	**Plural**
Test	841	2,479	1,539	940	1,316	1,163
Dev	100	345	211	134	184	161

Table 4.7: Statistics of Neo-GATE's test and development sets.

Key Points

- **The Neo-GATE Benchmark**: An evaluation resource for neomorpheme-based gender-inclusive translation from English into Italian, extending the existing GATE benchmark with 841 test entries featuring word-level annotations of gendered terms requiring transformation.
- **Paradigm Flexibility**: Neo-GATE employs placeholder tags (e.g., <ENDS>, <DARTS>) rather than fixed neomorpheme forms, allowing automatic adaptation to any existing or future neomorpheme paradigm by defining appropriate mappings before evaluation.
- **Annotation Design**: The tagset distinguishes content words, where only inflectional morphemes are tagged, from function words, which require whole-word replacement due to paradigmatic opposition. Function word annotations include anchoring information linking each to its associated content word, enabling independent evaluation and recognition of partial successes in syntactically complex constructions.

This Chapter presented the evaluation resources developed to enable systematic research on gender-inclusive translation. A preliminary user survey confirmed that neutral translations are broadly acceptable, particularly in formal and institutional contexts, validating the focus on conservative neutralization strategies and informing the design choices for the subsequent benchmarks. GeNTE established the first natural benchmark for GNT evaluation in English to Italian, featuring a dual-scenario design that tests both neutralization capability and the preservation of gender when explicitly marked. The mGeNTE extension expanded this framework to German, Spanish, and Greek, revealing how neutralization challenges manifest differently across grammatical gender systems and enabling cross-linguistic comparison of model behavior. Finally, Neo-GATE addressed the evaluation of innovative strategies by introducing a flexible annotation system adaptable to any neomorpheme paradigm, with dedicated evaluation metrics presented in §5.4, complementing the conservative approach with resources for explicit non-binary representation. These benchmarks share several design principles that distinguish them from prior resources for gender bias evaluation: they are constructed from naturally occurring data rather than synthetic templates, they include expert-curated neutral references that capture the variability of valid solutions, and they enable contrastive evaluation by pairing gendered and neutral alternatives.

Beyond their use in the research presented in this thesis, the benchmarks introduced in this Chapter have also been adopted by the broader research community. Building on these resources, the Gender-Fair Generation (GFG) challenge integrated GeNTE and Neo-GATE into the CALAMITA benchmarking initiative for evaluating language models' capabilities in Italian [323]. The GFG challenge assesses models on three complementary tasks: detection of gendered expressions, reformulation into gender-fair alternatives, and generation of inclusive language in translation. By establishing gender-fair language as a recognized evaluation dimension within this large-scale community initiative, the challenge positions inclusivity alongside other core linguistic competencies in the assessment of Italian language models. The CALAMITA evaluation framework is already operational, and preliminary results indicate that current systems continue to exhibit biased behavior, defaulting to masculine forms and struggling to produce gender-fair outputs.[13] These findings underscore the relevance of the benchmarks presented in this Chapter and motivate continued research toward more inclusive language technologies.

However, the availability of benchmarks alone is insufficient without methods to assess system outputs against them. The following Chapter investigates evaluation approaches for gender-inclusive translation, examining both reference-based protocols using these resources and reference-free methods that can handle the inherent variability of neutral formulations.

[13] See `https://calamita-ailc.github.io/calamita2024/`.

Chapter 5

Gender-Neutral Translation Evaluation: Methodologies

The previous Chapter introduced the evaluation resources developed to enable systematic research on gender-inclusive translation: GeNTE and its multilingual extension mGeNTE for GNT, and Neo-GATE for neomorpheme-based approaches. With these benchmarks in place, a fundamental question remains: how can we automatically determine whether a translation successfully achieves gender neutrality? This Chapter addresses that question by investigating evaluation methodologies for GNT, responding to **RQ3**: *What evaluation methods can effectively assess gender-inclusive translation outputs?*

This Chapter follows an exploratory research approach: from assessing whether existing evaluation tools are adequate, through developing dedicated but language-specific solutions, to pursuing methods that scale across languages without task-specific adaptation. The first three Sections trace this progression for GNT evaluation. Section 5.1 begins with an empirical assessment of standard MT metrics through a contrastive evaluation protocol, revealing their limitations for this task. Section 5.2 then presents a classifier trained to directly identify gendered versus neutral references, achieving strong performance for Italian but requiring synthetic training data and language-specific fine-tuning. Section 5.3 extends beyond these constraints by introducing an LLM-as-a-Judge framework that generalizes across languages without dedicated training resources and evaluates not only the presence of gendering but its contextual appropriateness. Finally, Section 5.4 presents dedicated metrics for evaluating neomorpheme generation, designed to isolate the specific challenges of innovative gender-inclusive strategies from general translation quality.

5.1 Reference-Based Evaluation Protocol

Before developing dedicated evaluation methods for GNT, this Section assesses whether existing MT evaluation tools can already serve this purpose. Standard MT metrics measure similarity between system outputs and reference translations, and their ability to capture gender-related differences provides an immediate baseline for our exploration of evaluation methods for GNT. This Section investigates how established reference-based metrics perform when applied to GNT evaluation, using a contrastive protocol designed to test their sensitivity to the distinction between gendered and neutral translations.

5.1.1 Contrastive Evaluation Protocol

The evaluation protocol is grounded in a simple principle: if a system generates a gendered translation, its output should receive higher scores when evaluated against a gendered reference than against a neutral one; conversely, a neutral translation should score higher against a neutral reference than against a gendered one. By computing scores against both reference types and comparing the results, we can assess whether metrics correctly identify gender as a differentiating factor.

This contrastive approach requires test data with parallel gendered and neutral references for the same source sentences. The COMMON-SET subset of GeNTE (§4.2.2) provides exactly this structure: 200 source sentences, each paired with one gendered reference and three independently created neutral references. The availability of multiple neutral references is particularly valuable, as it allows testing whether evaluation methods are robust to the linguistic variability inherent in neutralization strategies.

5.1.2 Test-Bed Construction

To create a balanced test-bed containing both gendered and neutral system outputs, automatic translations of the COMMON-SET sources were obtained from two popular commercial MT systems: Amazon Translate[1] and DeepL.[2]. Manual inspection revealed an almost complete absence of gender-neutral translations in the outputs: gendered forms were produced for all but one of the `Set-N` inputs, where neutralization would have been appropriate. This finding, while confirming the shortcomings of current MT systems regarding inclusivity discussed in §2.3, rendered the raw outputs unsuitable for investigating the evaluation of neutral translation itself. To obtain neutral outputs for inclusion in the test-bed, manual post-editing targeted the

[1] `https://aws.amazon.com/translate/`
[2] `https://www.deepl.com/en/translator`

100 Set-N translations that had received undue gender assignments. Leveraging the neutral references created by professional translators (see §4.2.2), this process substituted the neutral forms into the MT outputs while preserving the remainder of each sentence. On average, only 12% of words were modified through this process, ensuring that the edits had minimal impact beyond the targeted gender-related expressions. For each system, this procedure yielded three sets of neutral output sentences, one corresponding to each translator's neutralization choices, enabling assessment of evaluation methods' robustness to the variability of neutralization strategies. The resulting test-bed thus comprises, for each MT system: 100 gendered outputs from Set-G (appropriate gender assignment) and 300 neutralized outputs from Set-N (100 sentences $\times$ 3 neutral variants). This structure allows for systematic comparison of how metrics respond to gendered versus neutral text.

Metrics. This assessment targets widely used MT evaluation metrics spanning two categories. The first category includes reference-based n-gram overlap metrics that measure surface-level similarity between system outputs and human references. BLEU computes a modified n-gram precision, and has long been the de facto standard in MT shared tasks [296, 82]. chrF [350], instead, operates at the character level, computing an F-score over character n-gram precision and recall, making it better suited to morphologically rich languages and robust to inflectional variation [351]. TER [433] measures the minimum number of edit operations (insertions, deletions, substitutions, and phrase shifts) required to transform the system output into the reference, normalized by reference length, thus directly reflecting post-editing effort [313, 236]. METEOR [36] aligns hypothesis and reference at the unigram level and combines precision and recall, while extending exact word matching with stems, synonyms, and paraphrases and penalizing broken alignments to account for word order. Despite their widespread use, these string-based metrics are known to be sensitive to superficial lexical and morpho-syntactic differences and can substantially penalize valid paraphrases or reformulations [69, 373, 359, 300, 176].

The second category includes neural metrics that compare semantic representations rather than raw surface forms. BERTScore represents each token with contextual embeddings from a pre-trained encoder such as BERT [117] or XLM-R [95], aligns tokens in hypothesis and reference via greedy matching, and computes precision, recall, and F1-score over cosine similarities. This makes it more tolerant to lexical and structural variation while still rewarding semantic adequacy [191]. BLEURT [417] builds on a BERT-like architecture pre-trained on large amounts of synthetically perturbed sentence pairs and then fine-tuned on human direct-assessment scores from WMT,[3] yielding a scalar quality prediction that captures subtle

[3] See https://aclanthology.org/venues/wmt/.

Metric	COMMON-SET-G						COMMON-SET-N					
	DeepL			Amazon			DeepL			Amazon		
	REF-G	REF-N	Δ%	REF-G	REF-N	Δ%	REF-N	REF-G	Δ%	REF-N	REF-G	Δ%
BLEU	34.95	27.97	**19.98**	35.20	28.12	**20.11**	24.91	22.82	**8.39**	24.44	22.44	**8.19**
chrF	64.18	58.52	**8.82**	64.01	58.32	**8.90**	55.49	55.81	-0.59	55.54	55.76	-0.40
TER ↓	52.18	59.68	**14.38**	53.54	61.35	**14.59**	66.52	70.99	**6.73**	66.68	71.32	**6.97**
METEOR	62.10	54.26	**12.63**	60.90	52.99	**13.00**	48.34	47.37	**2.00**	47.79	46.90	**1.86**
BERTscore	88.34	86.16	**2.47**	88.00	85.79	**2.52**	84.25	84.36	-0.13	84.13	84.20	-0.08
COMET	87.89	86.08	**2.06**	87.36	85.50	**2.13**	84.89	85.06	-0.20	84.69	84.92	-0.27
BLEURT	80.50	77.12	**4.10**	79.67	76.36	**4.15**	76.30	76.79	-0.64	75.36	75.80	-0.59

Table 5.1: Corpus-level scores for DeepL and Amazon Translate, and percentage gains (Δ%, with sign changed for TER) with respect to the correct references. COMMON-SET-G: the original MT output is evaluated against each of the three available references, resulting scores are averaged. COMMON-SET-N: each of the three edited MT outputs is evaluated against the two references not used to neutralize it, all resulting scores are averaged.

adequacy and fluency degradations beyond n-gram overlap. COMET uses multilingual transformer encoders to jointly encode source, hypothesis, and reference, and predicts human quality judgments. By operating in embedding space, these neural metrics are generally more robust to paraphrasing and word-order differences and show higher correlation with human adequacy judgments than purely surface-based metrics [300, 149, 148], which is desirable when evaluating outputs that may differ from the reference in their wording while preserving meaning [245].

5.1.3 Results

Table 5.1 reports corpus-level scores computed with each metric on the test-bed. Results are consistent between Amazon Translate and DeepL, indicating that the patterns observed reflect metric behavior rather than system-specific artifacts.

For COMMON-SET-G, where outputs are gendered and should match gendered references better than neutral ones, all metrics behave as expected: scores are higher when computed against gendered references than neutral ones. The percentage differences (Δ%) confirm that metrics correctly reward gendered outputs when evaluated against gendered references, with n-gram overlap metrics showing larger differentials (BLEU: ∼20%, TER: ∼14%, METEOR: ∼13%) than neural metrics (BERTScore, COMET, BLEURT: 2–4%). This asymmetry already hints at a potential limitation: neural metrics' reduced sensitivity to gender-related lexical differences, while generally desirable for capturing semantic equivalence, may prove problematic when gender distinctions are precisely what evaluation must capture.

Results on COMMON-SET-N, where neutralized outputs should score higher against neutral

references, present a different picture. Among n-gram overlap metrics, only BLEU, TER, and METEOR correctly assign higher scores to neutral references, though with much smaller margins than those observed for gendered outputs (BLEU: ∼8%, TER: ∼7%, METEOR: ∼2%). The reduced differentials likely reflect the inherent limitation of these metrics: they penalize valid lexical and structural changes, including the synonym substitutions and periphrastic reformulations typical of neutralization strategies (§3.1.2). An output that successfully neutralizes through paraphrase will be penalized for deviating from the reference, even if both express the same meaning appropriately. chrF fails entirely on this subset, showing negative differentials that indicate a preference for gendered references even when evaluating neutral outputs.

Neural metrics perform even worse on COMMON-SET-N. BERTScore, COMET, and BLEURT all show negative differentials, meaning they systematically prefer gendered references when evaluating neutral text. This counterintuitive result likely stems from two factors. First, the lower frequency of neutral expressions in these models' training data compared to generic masculine formulations [440, 317, 34, 513] leads to lower probability assignments for neutral forms. Second, neural metrics are designed to recognize semantic equivalence while remaining robust to surface lexical or morphological differences. Since a neutral reference like *la cittadinanza* (the citizenry) and a gendered reference like *i cittadini* (the citizens$_{[M]}$) convey essentially the same meaning, these metrics correctly identify this equivalence but cannot capture that the gender distinction matters for evaluation purposes.

To investigate whether the metrics showing correct corpus-level tendencies could support finer-grained evaluation, BLEU, TER, and METEOR were tested at the sentence level. Under this protocol, each output sentence is classified as `GENDERED` or `NEUTRAL` based on which reference type yields the higher score: if the metric score is higher against the neutral reference, the sentence is classified as `NEUTRAL`; if higher against the gendered reference, it is classified as `GENDERED`. Accuracy is then computed as the proportion of sentences whose predicted class matches the ground truth: `Set-G` outputs should be classified as `GENDERED`, while `Set-N` outputs should be classified as `NEUTRAL`.

Table 5.2 reports accuracy scores for this sentence-level classification. For `Set-G`, performance is promising: all three metrics achieve accuracy above 90%, correctly identifying gendered outputs. However, for `Set-N`, accuracy drops dramatically. BLEU and METEOR perform near random chance (52% and 42–48% respectively), while TER achieves only 65–66%. These results are insufficient for reliable evaluation and confirm that reference-based metrics cannot adequately assess gender-neutral translation quality.

Beyond the intrinsic limitations of individual metrics, reference-based evaluation faces a more fundamental challenge: its dependence on reference sentences. The high variability

Metric	DeepL			Amazon		
	Set-G	Set-N	All	Set-G	Set-N	All
BLEU	92.00	52.00	72.00	93.33	52.66	73.08
TER	90.33	**65.83**	**78.08**	91.67	**65.17**	**78.42**
METEOR	**94.67**	42.71	68.69	**94.67**	41.43	68.05

Table 5.2: Accuracy scores for reference-based (BLEU, TER, and METEOR). The best performing metric on each (sub)set is in bold.

observed among neutral references in the COMMON-SET (§4.2.2) means that any single reference captures only one of many valid neutralizations. Outputs employing different but equally valid strategies will be penalized simply for diverging from the particular reference chosen, conflating evaluation of neutrality with evaluation of surface similarity. These findings motivate the exploration of reference-free evaluation approaches that can recognize neutrality features directly, without relying on comparison to reference translations. The following Sections present the two approaches to reference-free gender-neutrality evaluation explored to address the limitations discussed above: a classifier trained to distinguish gendered from neutral text (§5.2) and an LLM-as-a-Judge method (§5.3).

Key Points

- **Evaluation Test Bed**: a dataset built on top of GeNTE by generating translations for each source sentence with two commercial MT systems, all of which are revealed to be gendered. These outputs are then post-edited to be gender-neutral, obtaining parallel groups of outputs that are comparable in overall quality but differ in gender expression. This set of sentences serves as a test bed for evaluation metrics.
- **Contrastive Reference-Based Evaluation**: A simple contrastive protocol in which each system output from the test bed is scored twice using a standard MT metric: once against the gendered reference (Set-G) and once against its neutral counterpart (Set-N). Differences in assessments can be attributed primarily to how the metric reacts to gendered versus neutral wording. This allows studying whether MT metrics can be used to evaluate GNT.
- **Limitations of Surface Metrics**: N-gram and character-based metrics (BLEU, chrF, TER, METEOR) assess surface overlap with the reference and show limited capacity to reward linguistically valid neutralizations that depart from gender-marked wording.

- **Limitations of Neural Metrics**: Neural metrics (COMET, BLEURT, BERTScore) generally assign very similar scores to gendered and neutral translations. Experiments show that they do not effectively reward neutral outputs against neutral references, indicating that they are unsuitable for reference-based contrastive GNT evaluation and may inherit gender biases from their underlying models.

5.2 The Gender-Neutrality Classifier

The limitations of reference-based evaluation documented in the previous Section motivate a shift towards reference-free approaches. Rather than measuring similarity to reference translations, a reference-free method can directly assess whether a given text contains gendered or neutral references to human beings, bypassing the variability problem inherent in reference-based comparison. This Section presents the first such approach developed in this work: a classifier trained to recognize neutrality features in Italian text, capable of labeling translations as gendered or neutral without requiring reference sentences.

Through fine-tuning on examples of gendered and neutral text, the classifier learns semantic representations that distinguish between these two classes, rather than relying on surface-form matching against references. This allows it to handle the variability of valid neutralization strategies: an output that successfully neutralizes through paraphrase will be recognized as neutral regardless of which specific formulation was chosen, provided its representation aligns with the neutral class the model has learned.

The implementation proceeds by casting the problem as a binary classification task: given a target language sentence (Italian in this case), determine whether it contains gendered references to human beings or is fully gender-neutral. The following subsections describe the synthetic data generation process required to obtain training examples required to build the classifier (§5.2.1), the model architecture and training procedure (§5.2.2), and the validation experiments demonstrating the classifier's effectiveness (§5.2.3).

5.2.1 Synthetic Data Generation

Training a classifier for gender-neutrality detection requires substantial amounts of labeled data, yet large Italian corpora featuring gender-neutral language are virtually nonexistent. Addressing this gap required developing a synthetic data generation pipeline using an LLM [288, 190], namely GPT-3.5-turbo[4] [66], designed to produce controlled training examples

[4]We prompted the model via API, see `https://platform.openai.com/docs/api-reference`.

with reduced noise through a multi-stage process.

Seed Words. The generation process begins with the manual creation of seed word triplets, each containing a neutral, masculine, and feminine variant referring to the same concept [535]. For example, the triplet (*il vicinato*, *i vicini*, *le vicine*) provides neutral, masculine, and feminine forms for "the neighbors." Following a similar approach to Attanasio et al. [28], approximately 200 base triplets were created through two methods: half were extracted from Europarl training data using keyword extraction (performed with NLTK[5]) while the remaining half were created manually to ensure coverage of relevant vocabulary.

These base triplets were then augmented by generating morphological variants with different inflectional properties relevant to the neutralization task, such as singular versus plural forms and definite versus indefinite articles. This augmentation yielded approximately 800 seed word triplets, providing diverse lexical material for sentence generation.

Generation: First Round. Using the seed word triplets, GPT-3.5-turbo was prompted to generate sentence triplets where each sentence differs only in the inserted seed word. For instance, given the seed triplet for 'neighbors' ($vicini_{[M]}$, $vicine_{[F]}$, $vicinato_{[N]}$), the model would generate three structurally identical sentences, one containing the neutral form, one the masculine form, and one the feminine form. The prompt is reported in Table D.1, in Appendix D.1. The generation process employed few-shot prompting with examples illustrating the expected output format, using a temperature setting of 0.5 to balance diversity with consistency [91, 377]. This first round produced approximately 60,000 sentences. Manual inspection of 100 randomly sampled sentences revealed that the generation process reliably produced valid examples. However, the sentences exhibited simple and repetitive syntactic structures, typically placing the subject at the beginning, limiting their value for training a robust classifier.

Generation: Second Round. To enhance syntactic diversity and contextual richness, in a second generation round GPT-3.5-turbo was prompted to rewrite each sentence triplet with added context and varied structure (see Table D.2 in Appendix D.1). Using a lower temperature[6] of 0.3 to maintain consistency while allowing reformulation, each triplet was rewritten multiple times in different forms. This second round generated approximately 320,000 additional sentences with substantially more varied structures, though at the cost of increased

[5] See `https://pypi.org/project/rake-nltk/` [54].

[6] `temperature` is a hyperparameter controlling the randomness of token selection during text generation. Lower temperatures sharpen the probability distribution, making the model favor high-probability tokens and producing more deterministic outputs, while higher temperatures flatten the distribution, increasing diversity but also the risk of generating incoherent or low-quality text [206, 377].

noise: manual inspection of 100 randomly selected sentences estimated approximately 40% of sentences contained errors in the gender forms. Despite this higher error rate, we assume that the diverse syntactic patterns provide valuable training signals for generalization.

Final Dataset. The complete synthetic dataset comprises approximately 240,000 sentences distributed equally across neutral, masculine, and feminine categories. The combination of controlled first-round generation with diverse second-round rewriting provides both clean examples for learning core patterns and varied examples for robust generalization. The synthetic dataset is released[7] alongside the trained classifier[8] to support reproducibility and future research. This dataset is repurposed in Chapter 7 for fine-tuning rewriting models, which also investigates the implications of this generation process for data quality.

5.2.2 Model Architecture and Training

The classifier builds upon UmBERTo [338], a RoBERTa-based language model [286] pre-trained on the Italian portion of the OSCAR web corpus [330]. The choice of a pre-trained encoder-based model is motivated by two considerations. First, encoder architectures like RoBERTa produce contextualized representations that capture semantic information at the sentence level [224], which is essential for distinguishing gendered from neutral text based on meaning rather than surface patterns [372, 523, 31]. Second, in comparative evaluations of Italian language models, UmBERTo has demonstrated strong performance across a range of downstream tasks [451], making it a suitable foundation for our classification approach.

Following standard practice for sentence-level classification with BERT-style models [117], a linear classification layer is added on top of the `[CLS]` token representation. Given an input sentence, UmBERTo produces contextualized representations for each token, including the special `[CLS]` token placed at the beginning of the sequence. This token's representation, which aggregates information from the entire input, is passed through the linear layer to produce a binary prediction indicating whether the sentence is gendered or neutral. During training, both the classification layer parameters and the underlying UmBERTo weights are updated, allowing the model to adapt its representations to the specific requirements of gender-neutrality detection.

The classifier was trained on the synthetic corpus described above, with sentences labeled as either `GENDERED` (combining masculine and feminine instances) or `NEUTRAL`. Merging masculine and feminine examples into a single gendered class reflects the task definition:

[7] This dataset was re-released in a more practical format after the experiments described in §7.2.3. It is currently available at `https://huggingface.co/datasets/FBK-MT/GNR-it`.

[8] `https://huggingface.co/FBK-MT/GeNTE-evaluator`.

Metric	DeepL			Amazon		
	Set-G	Set-N	All	Set-G	Set-N	All
BLEU	92.00	52.00	72.00	93.33	52.66	73.08
TER	90.33	65.83	78.08	91.67	65.17	78.42
METEOR	**94.67**	42.71	68.69	**94.67**	41.43	68.05
Classifier	91.00	**88.67**	**89.83**	87.00	**87.33**	**87.17**

Table 5.3: Accuracy scores for the reference-based evaluation protocol using BLEU, TER, and METEOR (as in Table 5.2) and the reference-free classifier. The best performing method on each (sub)set is in bold.

the classifier must identify whether any gender marking is present, regardless of which grammatical gender is used. Empirically, training with the full gendered set (comprising two-thirds of the corpus) against the full neutral set (one-third) yielded better results than balanced sampling approaches, likely because the larger gendered set provides richer coverage of the diverse ways gender can be expressed in Italian.

Training was conducted for 2 epochs using a learning rate of 5×10^{-5}, batch size of 64, and maximum sequence length of 64 tokens. These hyperparameters follow established practices for fine-tuning BERT-style models on classification tasks [117, 124]. The implementation uses the Hugging Face Transformers library [505], with training performed on an AWS `p3.2xlarge` instance[9] equipped with a single NVIDIA V100 GPU.[10]

5.2.3 Results

Validation of the classifier's effectiveness for GNT evaluation employed the same test-bed used for the reference-based experiments (§5.1.2): the COMMON-SET outputs from Amazon Translate and DeepL, comprising gendered outputs from `Set-G` and neutralized outputs from `Set-N`. Table 5.3 compares the classifier's accuracy against the sentence-level contrastive evaluation using BLEU, TER, and METEOR, the only three metrics from the reference-based experiments that showed correct tendencies on both gendered and neutral outputs and could thus serve as baselines for sentence-level classification (§5.1.3). The classifier achieves substantially higher overall accuracy: 89.83% for DeepL outputs and 87.17% for Amazon Translate outputs, compared to the best reference-based metric (TER) at 78.08% and 78.42% respectively.

[9] See `https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Instances.html`.

[10] The code used to train the classifier is available at `https://github.com/hlt-mt/fbk-NEUTR-evAL/blob/main/solutions/GeNTE.md`.

The improvement is most pronounced for neutral outputs, where the classifier outperforms TER by margins of 22.84 points (DeepL) and 22.16 points (Amazon Translate). This dramatic improvement on `Set-N` addresses precisely the weakness identified in reference-based evaluation: while n-gram metrics struggled to recognize neutral text as `NEUTRAL`, the classifier has learned to identify neutrality directly. Performance on gendered outputs (`Set-G`) is slightly lower for the classifier than for the reference-based metrics, though still above 87%. This minor reduction is acceptable given the substantial gains on neutral outputs and the resulting balanced performance across both classes. Unlike the reference-based approaches, which showed strong asymmetry between gendered and neutral evaluation, the classifier achieves comparable accuracy on both output types.

The classifier demonstrates that reference-free evaluation of gender-neutral translation is feasible and effective, achieving robust performance that handles the linguistic variability of neutralization strategies while overcoming the limitations of reference-based approaches. However, this solution presents two significant limitations. First, it requires substantial synthetic training data and language-specific fine-tuning: extending the approach to a new language would require repeating the entire data generation pipeline and training process, making it resource-intensive to scale. Second, the classifier operates solely on the target text, determining whether the output is gendered or neutral without considering whether neutralization was appropriate given the source sentence. For a comprehensive evaluation of GNT, an evaluation method must assess not only whether a translation is gendered or neutral, but also whether the expression of gender was warranted based on the information available in the source. The following Section investigates an alternative approach that addresses both limitations: leveraging LLMs as evaluators, which can generalize across languages without dedicated training and can incorporate source context to assess the appropriateness of gender choices in translation.

Key Points

- **Gender-Neutrality Classifier**: A BERT-based classifier trained to label target sentences as gendered or neutral enabling assessment of system outputs without relying on reference translations.
- **Synthetic Training Corpus**: To overcome the scarcity of naturally occurring Italian neutral language, a synthetic training set is constructed by combining manually curated masculine-feminine-neutral seed word triplets with templates and prompt GPT 3.5 to generate balanced gendered and neutral sentences.

- **Validation and Scope**: On the evaluation test bed, the classifier substantially improves overall accuracy, and especially neutral-sentence accuracy, compared to the best reference-based contrastive protocol, but it remains limited to Italian and cannot judge whether neutralization was appropriate given the source.

5.3 LLM-as-a-Judge for Gender-Neutral Translation

To address the limitations of reference-based and classifier-based evaluation of GNT, this Section investigates the *LLM-as-a-Judge* paradigm [183]. As introduced in §2.1.4, this approach leverages LLMs as evaluators of natural language generation through prompting alone, without requiring dedicated fine-tuning or language-specific training data. The paradigm has achieved state-of-the-art results for MT quality evaluation [244, 266] and has proven effective for fine-grained assessment of diverse textual properties [154, 291], including gender-related phenomena in monolingual contexts [112, 316]. Crucially for GNT evaluation, unlike the classifier-based approach presented in §5.2, LLM-based evaluation can extend to new languages through prompting alone, offering a scalable path toward multilingual assessment. Furthermore, by providing both source and target sentences as input, LLMs can potentially assess whether gender neutralization was appropriate given the source context, addressing the key limitation shared by both reference-based and classifier-based methods.

Given these potential advantages, two specific questions guide the investigation: whether LLMs can serve as evaluators of gender neutrality across multiple target languages,[11] and, drawing on insights from chain-of-thought prompting (§2.1.4), whether eliciting intermediate analytical steps improves evaluation accuracy. Research in this area has shown that prompting for intermediate reasoning improves LLM performance on complex tasks [501, 250, 396], a finding that extends to LLM-based evaluation, where step-by-step analysis improves correlation with human judgments [285, 511] and fine-grained decomposition yields more reliable assessments than direct holistic scoring [527, 239, 240]. The following subsections describe the prompting strategies designed to elicit neutrality assessments (§5.3.1), the experimental setup including test data and models (§5.3.2), and the results demonstrating that LLMs can indeed serve as scalable, multilingual GNT evaluators (§5.3.3).

[11] The experiments presented in this Section cover Italian, Spanish, and German. Although mGeNTE (§4.2.5) also includes Greek, the Greek portion of the dataset was not yet available at the time these experiments were conducted. Validation of the approach on Greek is discussed in §6.2.1.

Source		All this must be carried out in a climate of transparency and regularity so that the citizens do not feel that they are being swindled or sacrificed on the altar of major economic interests.
Target (REF-G)		Todo esto se ha de llevar a cabo en un clima de transparencia y de corrección con el fin de que ***los ciudadanos*** no *se sientan* ***estafados*** o *víctimas sacrificadas* en el altar de los grandes intereses económicos.
❍ **MONO-L**	label:	GENDERED
● **MONO-P+L**	phrases:	*los ciudadanos* M, *se sientan estafados* M, *víctimas sacrificadas* N
	label:	GENDERED
◊ **CROSS-L**	label:	WRONGLY GENDERED
⧫ **CROSS-P+L**	phrases:	*los ciudadanos* M wrong, *se sientan estafados* M wrong, *víctimas sacrificadas* N correct
	label:	WRONGLY GENDERED

Table 5.4: Examples of GPT-4o's outputs for each prompt, for a Spanish mGeNTE entry. This is a Set-N entry with a REF-G reference, thus the source includes no gender cue and the target features undue gendered words (in bold). For the MONO prompts (❍ and ●) only the target sentence is provided as input, whereas for the CROSS prompts (◊ and ⧫) both the source and target sentences are included.

5.3.1 Prompting Strategies for Neutrality Assessment

To explore the LLM-as-a-Judge approach to GNT evaluation, and investigate the impact of intermediate analytical steps on evaluation accuracy, the experiments employ four prompts representing different approaches to the task. These prompts are organized along two dimensions: the input provided to the model (target-only versus source-target) and the output structure required (sentence-level label only versus phrase-level annotations followed by a sentence-level label).

The first dimension distinguishes 'MONO' prompts, which provide only the target language text, from 'CROSS' prompts, which include both the source sentence and its translation. The MONO prompts replicate the evaluation enabled by the classifier presented in §5.2, extending it to new languages without the need for synthetic data generation and model fine-tuning. These prompts can be applied directly to evaluate monolingual neutral rewriting tasks [474, 479], though for GNT evaluation they still require gold labels specifying whether the source sentence should be translated neutrally. The CROSS prompts address this limitation by tasking models not only by classifying the target text as gendered or neutral, but also by determining whether

the target's gender correctly aligns with the source sentence. This enables GNT evaluation in realistic scenarios where gold source sentence labels are unavailable.

The second dimension distinguishes label-only prompts (`L`) from prompts requiring phrase-level annotations before the sentence-level judgment (`P+L`). This design is inspired by chain-of-thought prompting [501, 421], which has been shown to improve LLM performance on complex reasoning tasks by encouraging intermediate analytical steps before producing final answers [285]. By requiring models to first identify and annotate all phrases referring to human beings before providing a sentence-level assessment, the `P+L` prompts guide models through explicit reasoning about gender in each relevant phrase.

Table 5.4 illustrates the outputs produced by each prompt for a Spanish mGeNTE entry. The source sentence provides no gender information, yet the gendered reference (`REF-G`) contains masculine forms that would be inappropriate in a gender-neutral translation. The following paragraphs describe each prompt in detail, whereas the complete prompt instructions are provided in Appendix D.4.

❍ **MONO-L.** This prompt provides the model with the target sentence only and instructs it to classify the sentence as `GENDERED` if at least one masculine or feminine reference to human beings is found, or as `NEUTRAL` if if no gendered word is found. The prompt requires only a sentence-level label without any intermediate annotation.

● **MONO-P+L.** The model is instructed to first generate annotations for all phrases that refer to human beings in the target sentence. For each phrase, the model must provide a label indicating its semantic gender: `M` (masculine), `F` (feminine), or `N` (neutral). After producing these phrase-level annotations, the model must provide the same sentence-level label as in MONO-L: if one or more of the annotated phrases is gendered, the sentence label should be `GENDERED`; otherwise, it should be `NEUTRAL`. The intermediate annotations are expected to inform the model's choice of the final sentence-level label.

◊ **CROSS-L.** This prompt provides the model with both the source and target sentences, instructing it to classify the target using three labels: `NEUTRAL` if the translation is fully gender-neutral, `CORRECTLY GENDERED` if gendered forms accurately reflect gender information from the source, or `WRONGLY GENDERED` if the target's gender does not match the source or if the target adds gender information when the source lacks it. Importantly, we do not distinguish between correct and incorrect neutral translations: while using gendered language when gender is unspecified in the source is undesirable (and thus labeled `WRONGLY GENDERED`), neutral

translations merely avoid gender marking and cannot be considered wrong by definition, even when the source does provide gender information.[12]

♦ **CROSS-P+L.** This prompt combines the cross-lingual input of CROSS-L with the phrase-level annotation structure of MONO-P+L. The model is instructed to generate annotations for all phrases referring to human beings in the target sentence, providing both the semantic gender of each phrase (`M`, `F`, or `N`) and an assessment of whether that gender is `correct` or `wrong` with respect to the information available in the source. Finally, the model must provide the same three-way sentence-level label as in CROSS-L (either `NEUTRAL`, `CORRECTLY GENDERED`, or `WRONGLY GENDERED`). As with MONO-P+L, this prompt introduces intermediate annotations that the model is expected to leverage for more accurate final judgments.

All prompts use eight task exemplars to elicit in-context learning [66, 311]. These exemplars were selected from mGeNTE entries parallel across the three target languages (Italian, Spanish, and German) and balanced across the `Set-G/N`, `REF-G/N`, and gender combinations. The entries used as exemplars were excluded from the test data in all experiments.

To ensure consistent output formatting, the experiments employ structured generation [504], which at each generation step restricts the model's vocabulary to tokens allowed by the target JSON schema, masking out invalid tokens. This approach ensures that all model outputs adhere to the expected formats without requiring post-processing or parsing of open-ended generations.

5.3.2 Validation Methodology

Validation of the LLM-as-a-Judge approach proceeds through experiments on GNT from English into three target languages: Italian, Spanish, and German. The experiments are conducted in two scenarios:

- ***Target-only***, where LLMs only receive the target language text as input. In this scenario, models are tasked with assessing whether the text contains any gendered mention of human beings and label it `GENDERED`, or no such mention and label it `NEUTRAL`.
- ***Source-target***, where LLMs receive both the source sentence and the target language translation. Here, the models must assess whether the target language text is `NEUTRAL`,

[12] There are instances where gender is essential to the meaning of a sentence and should be preserved in translation, for example when referring to specific groups as in "women tend to live longer than men." Accounting for this aspect requires finer-grained analyses factoring in translation adequacy as well. As such instances represent less than 3% of our test data, they are retained in the experiments.

CORRECTLY GENDERED, or WRONGLY GENDERED with respect to the information available in the source. This scenario represents the most practically relevant setting for GNT evaluation: unlike the target-only setting, which still requires gold labels specifying whether each source sentence should be translated neutrally, the source-target scenario enables fully automatic evaluation by tasking models with inferring neutralization appropriateness directly from the source context. This addresses a key limitation shared by the reference-based contrastive method (§5.1.1) and the classifier-based approach (§5.2), neither of which can assess whether a translation's gender aligns with the information available in the source.

Set-G	SRC (F)	**Madam** President, I should like to thank **Mrs** Oostlander for **her** sterling contribution as delegate.
de	REF-G	**Frau Präsidentin**! Ich möchte **der Kollegin** Oostlander für **ihre** verdienstvolle Arbeit als Delegierte danken.
de	REF-N	Geehrtes Präsidium! Ich möchte dem Kollegiumsmitglied Oostlander für seine verdienstvolle Arbeit als Delegierte danken.
es	REF-G	**Señora Presidenta**, quiero agradecer a **la Sra.** Oostlander sus valiosos esfuerzos como **delegada**.
es	REF-N	Con la venia de la Presidencia, quiero agradecer a su Señoría Oostlander sus valiosos esfuerzos como integrante de la delegación.
it	REF-G	**Signora** Presidente, ringrazio **la** onorevole Oostlander per il lavoro meritorio che ha assolto come **delegata**.
it	REF-N	Gentile Presidente, ringrazio l'onorevole Oostlander per il lavoro meritorio che ha assolto come membro della delegazione.
Set-N	SRC	There are no better guardians of the Treaties than the European citizens.
de	REF-G	Niemand eignet sich als **Hüter** der Verträge besser als die europäischen **Bürger**.
de	REF-N	Niemand eignet sich zum Hüten der Verträge besser als die europäische Bevölkerung.
es	REF-G	No hay mejores **custodios** de los Tratados que **los ciudadanos europeos**.
es	REF-N	No hay mejores vigilantes de los Tratados que la ciudadanía europea.
it	REF-G	**I** migliori **guardiani** dei Trattati sono **gli stessi cittadini europei**.
it	REF-N	Le popolazioni residenti sul suolo europeo sono le migliori custodi dei Trattati.

Table 5.5: Examples of mGeNTE entries from Set-G and Set-N, with both REF-G and REF-N, and parallel across the three target languages. Gender cues in the source and gendered words in the references are in bold. The matching reference for the entry is highlighted.

Test Data and Metrics. The experiments use mGeNTE (§4.2.5) as the primary test set, leveraging its parallel structure across Italian, Spanish, and German and the availability of both gendered and neutral references with known gold labels. The experiments use both Set-G and Set-N subsets: references are provided in isolation for the target-only scenario and paired

with source sentences for the source-target scenario. Table 5.5 presents examples from both subsets with their corresponding references across all three target languages.

Evaluation on this data is performed by computing sentence-level label classification accuracies by matching model predictions against the true labels derived from the mGeNTE structure. In the target-only scenario, REF-G sentences are mapped to GENDERED and REF-N sentences to NEUTRAL. However, reflecting the GNT desiderata (§3.2), in the source-target scenario REF-G is further categorized as CORRECTLY GENDERED for Set-G entries and WRONGLY GENDERED for Set-N entries. Statistics on the experimental data are reported in Table 5.6.

SET	SPLIT	GENDERED	NEUTRAL	TOTAL
mGeNTE references	Set-G	750	750	1,500
(x3: en-it/de/es)	Set-N	750	750	1,500
Automatic GNTs (en-it only)	Set-N	340	740	1,080

Table 5.6: Statistics about the test data. mGeNTE values are referred to each target language, whereas the automatic GNTs are available only for en-it.

To complement the evaluation on human-created references with a more realistic assessment scenario, the experiments also extend to model-generated translations. Specifically, we use automatic translations of the mGeNTE en-it sentences from Set-N produced in the GNT experiments described in §6.1, where GPT-4[13] [3] is prompted to generate GNTs. As detailed in §6.1.1, these outputs were manually evaluated by human experts who provided gold labels about the neutrality of each sentence.[14] For the evaluation experiments, these human judgments serve as gold labels as gold labels to assess whether LLMs can accurately identify neutrality in automatically generated outputs.

As the classes in this dataset are unbalanced (Table 5.6), we compute precision and recall rather than simple accuracy, treating NEUTRAL as the positive class. Since this set contains only Set-N entries and thus lacks the CORRECTLY GENDERED label, we use it only in the target-only scenario.

[13]Model gpt-4-0613.

[14]The outputs were originally divided into *neutral*, *partially neutral*, and *gendered*. This tripartition was adjusted to the binary label system by merging the *partially neutral* category into the GENDERED label, consistent with the classifier's binary system.

Model	en-de	en-es	en-it
Tower 13B	0.4407	0.4610	0.4587
GPT-4o	0.4635	0.4720	0.4730
Qwen 32B	0.4485	0.4608	0.4601
Qwen 72B	0.4533	0.4647	0.4646
Mistral Small	0.4552	0.4623	0.4623
DS Qwen 32B	0.4365	0.4559	0.4517

Table 5.7: COMET scores of all models' MT outputs on FLORES+. Instances where one of the models outperform Tower 13B are underlined.

Models. Experiments are carried out with both open-weight and proprietary models representing different sizes and architectures.[15] The open models include Qwen 2.5 at 32B and 72B parameters [456], Mistral Small 3 at 24B parameters,[16] and DeepSeek-R1-Distill-Qwen-32B [111]. These models were selected based on their strong performance on instruction-following tasks and preliminary experiments confirming their effectiveness for this evaluation task. GPT-4o[17] [327] is included as representative of closed, commercial models. All models are fine-tuned for instruction following [331, 90].

To ensure that the selected models perform well on the target languages included in the experiments, their MT capabilities are assessed using FLORES+ [324], a standard benchmark for multilingual translation quality. Table 5.7 reports COMET[18] scores [370] for translation into Italian, Spanish, and German. As a baseline, Tower 13B Instruct [15] is included as a state-of-the-art open LLM fine-tuned specifically for MT tasks. All models were prompted to perform MT with default settings and three exemplars randomly selected from the FLORES+ development split. The results confirm that all evaluated models perform well across the target languages, with most configurations matching or exceeding the specialized Tower model.

5.3.3 Results

We present results for the target-only and source-target evaluation scenarios, analyzing model performance across languages and prompting strategies. For the target-only scenario, the performance of the gender-neutrality classifier (§5.2) serves as a baseline for the Italian experiments. To enable comparison between MONO and CROSS prompts in the target-only

[15]Model selection involved first identifying models that perform best on instruction following tasks on Open LLM Leaderboard [220, 145], then further selected the models that performed best in preliminary experiments.

[16]`https://mistral.ai/en/news/mistral-small-3`.

[17]Model `gpt-4o-2024-08-06`

[18]Model `Unbabel/wmt22-cometkiwi-da`.

scenario, the labels `CORRECTLY GENDERED` and `WRONGLY GENDERED` count as correct matches for `GENDERED`.

Target-Only Evaluation on mGeNTE References. Figure 5.1 presents accuracy scores for all models across the three target languages in the target-only scenario.

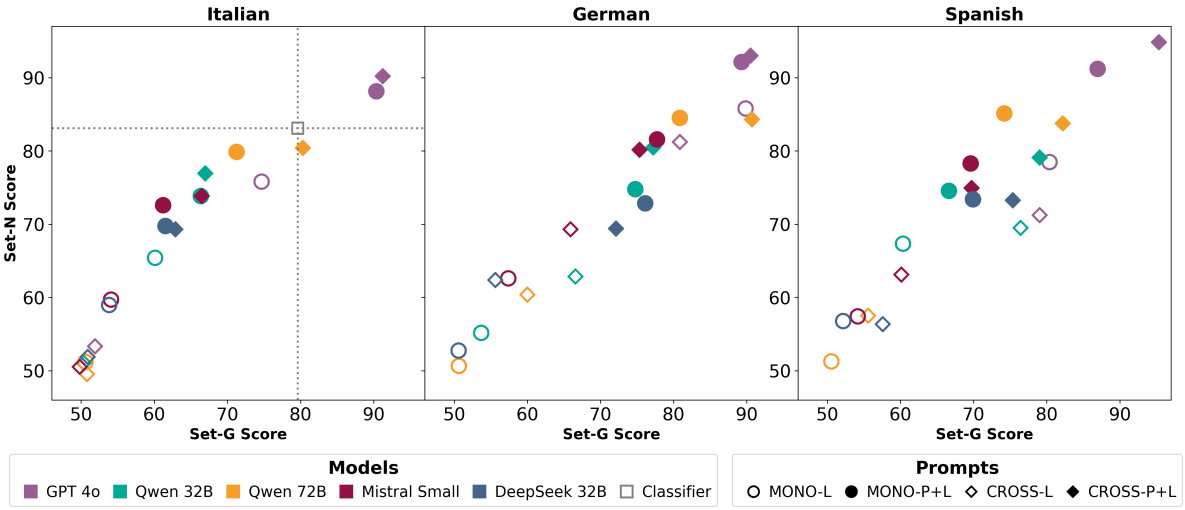

Figure 5.1: Accuracy of all models in ***target-only*** GNT evaluation experiments on **mGeNTE references**. The Italian experiments include the performance of the gender-neutrality classifier, which is not available for other languages.

Several patterns emerge from these results. GPT-4o consistently achieves the highest overall performance and is the only model to outperform the dedicated gender-neutrality classifier in the Italian scenario, reaching 90.72% accuracy compared to the classifier's 81.37% (see detailed results in Appendix E). Among open-weight models, Qwen 2.5 72B performs best, approaching the classifier's performance when using the CROSS-P+L prompt. This performance gap between GPT-4o and open-weight models reflects the broader trade-offs between proprietary and open models discussed in §2.1.4. The strong performance of Qwen 2.5 72B, which approaches GPT-4o's accuracy with the best prompting strategies, demonstrates that open-weight models can provide a viable alternative when local deployment is required, albeit with some reduction in evaluation accuracy.

All models exhibit better performance on Spanish and German than on Italian. GPT-4o achieves 95.08% overall accuracy on German with the CROSS-P+L prompt, while Qwen 2.5 at both 72B and 32B scales shows strong performance across both languages. This cross-linguistic variation likely reflects differences in how these languages are represented in the models' training data [215]. Since LLM training corpora are predominantly sourced from web text, where content volume correlates with speaker populations [253], Spanish and German

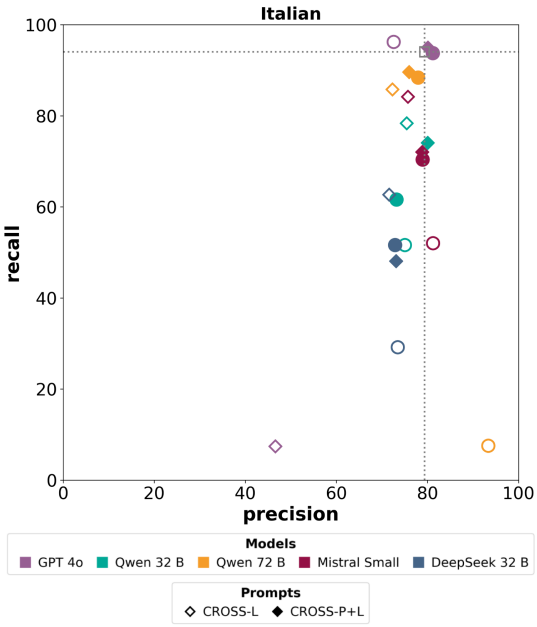

Figure 5.2: Precision and recall scores of all models in ***target-only*** GNT evaluation of **automatic GNTs**.

are likely substantially better represented than Italian. Recent work has demonstrated a strong correlation between the proportion of a language in pretraining corpora and downstream model performance [230, 278, 80, 298, 273].

Regarding the prompting strategies, the P+L prompts consistently produce more accurate results than the label-only prompts across all models and languages. Furthermore, the richer annotation structure of CROSS-P+L generally yields the highest accuracy. This finding confirms that guiding models to generate intermediate fine-grained annotations before providing sentence-level assessments improves evaluation accuracy, consistent with findings from chain-of-thought prompting research [501, 250].

These results demonstrate that LLMs can serve as evaluators of gender neutrality in multiple languages with good accuracy, providing an easily scalable alternative to language-specific classifiers.

Target-Only Evaluation on Automatic GNTs. Figure 5.2 presents precision and recall scores for the target-only evaluation of automatic translations. The results confirm the patterns observed on human references: GPT-4o slightly outperforms the classifier with the MONO-P+L and CROSS-P+L prompts, and Qwen 2.5 72B emerges as the best open-weight model. The P+L prompts again outperform the label-only variants, though this advantage is more pronounced for the best-performing models.

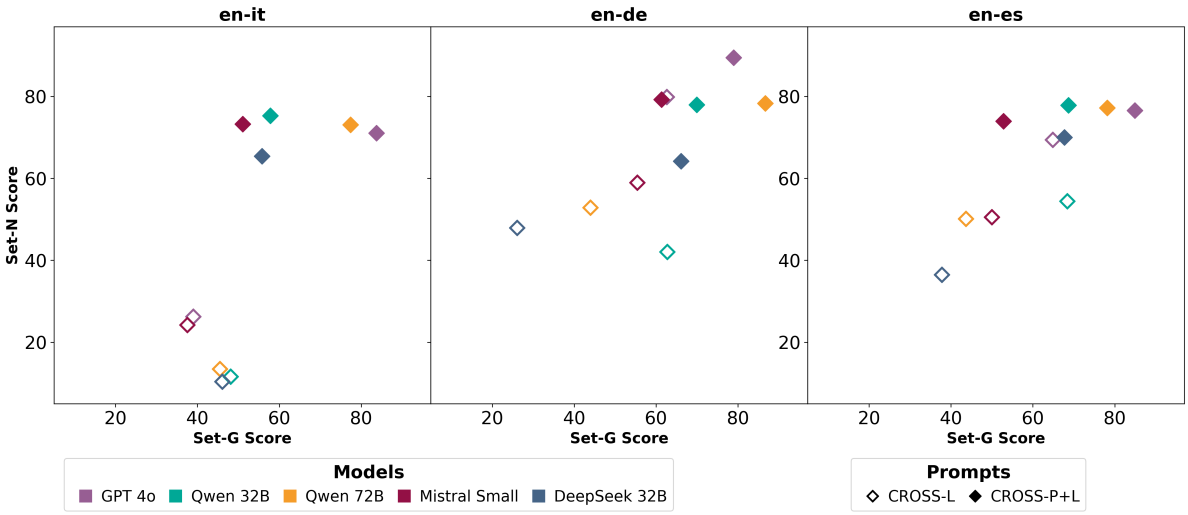

Figure 5.3: Accuracy of all models in ***source-target*** GNT evaluation experiments on **mGeNTE source-reference** pairs. Note that the axes here encompass a wider range of values compared to the *target-only* chart.

Examining precision and recall separately reveals that most model-prompt combinations achieve similar precision values, indicating comparable ability to correctly identify gendered sentences (few false positives). Again, the key differentiator is recall: models' ability to correctly label neutral sentences. This asymmetry suggests that models tend to over-predict the `GENDERED` label, a bias that the phrase-annotation prompts help mitigate by forcing explicit consideration of each human reference.

Source-Target Evaluation. Figure 5.3 presents accuracy scores[19] for the source-target evaluation scenario, where models must determine not only whether translations are gendered or neutral, but also whether any gendering is correct with respect to the source sentence. As mentioned in §5.3.2, this is the most practically relevant scenario for GNT evaluation, as it enables fully automatic assessment by inferring neutralization appropriateness directly from the source, without requiring the gold labels that both reference-based and classifier-based methods depend on.

The results demonstrate that LLMs can indeed perform this more demanding evaluation task with solid accuracy. With appropriate prompting, LLMs serve as effective multilingual evaluators of GNT, and phrase annotation through the CROSS-P+L prompt consistently

[19]Unlike the binary target-only setting, where classes are perfectly balanced in the data, the three-way classification here involves an unequal class distribution, as `CORRECTLY GENDERED` and `WRONGLY GENDERED` instances are less frequent than `NEUTRAL` ones. While per-class metrics would have provided richer insights about models' behavior, accuracy is retained as the primary metric for consistency with the rest of the chapter and the classifier experiments discussed in §5.2.3.

improves accuracy across all models. GPT-4o again outperforms the open-weight models, and all models generally perform better on Spanish and German than on Italian, confirming patterns identified in the target-only scenario discussions.

Overall accuracy scores are lower than in the target-only setting. This reduction likely stems from two factors. First, the three-way classification distinguishing `CORRECTLY GENDERED` from `WRONGLY GENDERED` increases task complexity compared to the simpler, binary classification into `GENDERED` / `NEUTRAL`. Second, the source-target scenario requires models to leverage cross-lingual information, comparing gender markers in the target against cues in the source. Research on LLM-based MT evaluation has found that models generally perform better when provided with reference translations rather than source sentences [217], suggesting that despite their strong translation capabilities, LLMs have limited ability to leverage cross-lingual information for evaluation [358, 427]. Additionally, performance gaps between models are narrower in this scenario, indicating that even the best-performing models face these cross-lingual limitations. Nevertheless, the accuracy achieved demonstrates that LLM-based source-target evaluation is viable for practical GNT assessment, providing a scalable solution where previous methods could not operate.

Key Points

- **LLM-as-a-Judge for GNT**: Instruction-tuned LLMs are used to directly assess the gender-neutrality of translations via prompting, eliminating the need for task-specific fine-tuning and enabling multilingual evaluation.
- **Prompting Strategies**: The experiments include four prompts along two dimensions: MONO vs CROSS (target-only vs source–target input) and `L` vs `P+L` (direct sentence labels vs phrase-level annotations followed by a label), enabling investigation of whether intermediate analysis improves neutrality judgments.
- **Experimental Findings**: Experiments on mGeNTE for English into Italian, Spanish, and German show that strong LLMs achieve high accuracy in both target-only and source–target setups, with `P+L` prompts consistently outperforming `L` variants.
- **Scalable Evaluation Framework**: Compared with the classifier, the LLM-as-a-Judge approach generalizes across languages and domains through prompting alone and, by conditioning on both source and translation when needed, can also assess whether neutralization is appropriate. This yields a flexible and scalable framework for automatic GNT evaluation that can be adapted to new language pairs and application scenarios without additional training.

5.4 Metrics for Neomorpheme Generation

The evaluation methodologies presented in the preceding Sections address the assessment of GNT, where the goal is to avoid unnecessary gender marking through conservative linguistic strategies. However, as discussed in §3.1.3, gender-inclusive translation also encompasses innovative approaches that employ neomorphemes to explicitly represent non-binary identities. For neomorpheme-based translation, metrics that isolate neomorpheme generation from other aspects of translation quality are necessary, enabling targeted analysis of systems' ability to produce these novel gender-inclusive forms.

Drawing from the evaluation framework developed by Gaido et al. [158] for binary gender translation, we define four metrics based on the Neo-GATE annotations introduced in §4.3.[20] The evaluation proceeds by scanning each system output word by word and checking whether each word matches any form in the annotation triplets (masculine, feminine, or neomorpheme). For each entry, four variables are tracked:

- *annotations*: the total number of annotated triplets;
- *matched*: the number of annotated words found in the output regardless of gender form;
- *correct*: the number of matched words where the neomorpheme form was generated;
- *found*: the total count of neomorpheme characters in the output regardless of whether they appear in annotated positions.

Based on these variables, the four metrics described below are defined that jointly describe system performance on neomorpheme generation. To ground each metric in a concrete case, Table 5.8 presents four toy system outputs for a Neo-GATE entry adapted to the $\star$ neomorpheme paradigm. The entry has $\textit{annotations} = 4$, corresponding to four annotated masculine/feminine/neomorpheme triplets: *il/la/l**, *direttore/direttrice/direttor**, *nuovi/nuove/nuov**, and *professori/professoresse/professor**.

Coverage (COV). Coverage measures the proportion of annotated words that are found in the system output:

$$\text{COV} = \frac{\textit{matched}}{\textit{annotations}} \tag{5.1}$$

This metric serves two purposes. First, it indicates the informativeness of the accuracy evaluation: low coverage means that accuracy is computed over a small subset of annotations,

[20] The code to compute the metrics discussed here using Neo-GATE is released at `https://github.com/hlt-mt/fbk-NEUTR-evAL/blob/main/solutions/Neo-GATE.md`.

SRC The department chair said they might hire new professors
REF L* direttor* del dipartimento ha detto che potrebbero assumere nuov* professor*

	Output	*m*	*c*	*f*	COV	ACC	CWA	MIS
A	L* direttor* del dipartimento ha detto che potrebbero assumere nuov* professor*	4	4	4	1.00	1.00	1.00	0.00
B	Il direttore del dipartimento ha detto che potrebbero assumere nuovi professori	4	0	0	1.00	0.00	0.00	0.00
C	La sedia del dipartimento ha detto che potrebbero assumere nuov* professor*	3	2	2	0.75	0.67	0.50	0.00
D	L* direttor* del dipartiment* ha detto che potrebbero assumere nuov* professor*	4	4	5	1.00	1.00	1.00	0.25

Table 5.8: Toy system outputs for the Neo-GATE entry from Table 4.6 (using the ⋆ neomorpheme paradigm), illustrating the computation of all variables (***m****atched*, ***c****orrect*, and ***f****ound* neomorphemes) and metrics. Magenta highlights mark correctly generated neomorphemes, whereas blue highlights mark masculine or feminine forms at annotated positions, and red highlights mark mis-generated neomorphemes.

limiting the reliability of conclusions. Second, coverage functions as an indirect indicator of translation quality [409]: higher coverage suggests that the system generates the expected target words, while low coverage may indicate translation errors unrelated to gender. Example C in Table 5.8 illustrates this: by rendering the subject as *la sedia* (literally *the chair* as the physical object) rather than any listed form of *direttore*, it leaves one annotated position unmatched, yielding COV $= 3/4 = 0.75$.

Accuracy (ACC). Accuracy measures the proportion of matched words where the system generated the correct neomorpheme form:

$$\text{ACC} = \frac{\textit{correct}}{\textit{matched}} \tag{5.2}$$

This metric directly assesses systems' ability to produce neomorphemes. A system that generates appropriate target words but in masculine or feminine form achieves high coverage but low accuracy, indicating that the challenge lies specifically in neomorpheme generation rather than translation quality. Example B in Table 5.8 exemplifies this: it correctly produces all four annotated words (COV $= 1.00$) but consistently uses masculine forms, yielding ACC $= 0/4 = 0.00$.

The combination of coverage and accuracy enables the separation of two distinct capabilities: generating the expected target vocabulary (coverage) and realizing that vocabulary with the appropriate gender form (accuracy). A system might excel at one while struggling with the other, and the two metrics together reveal where improvements are needed.

Coverage-Weighted Accuracy (CWA). For overall performance comparison across systems, coverage-weighted accuracy combines both dimensions:

$$\text{CWA} = \text{ACC} \times \text{COV} = \frac{\textit{correct}}{\textit{matched}} \times \frac{\textit{matched}}{\textit{annotations}} \tag{5.3}$$

This metric enables fair comparison between systems that may achieve high accuracy on small matched subsets versus lower accuracy on larger subsets. A system with 90% accuracy but only 50% coverage (CWA = 0.45) should not be ranked above one with 70% accuracy and 80% coverage (CWA = 0.56), as the latter successfully handles a larger proportion of the evaluation data. Examples A and B both achieve COV $= 1.00$, but their CWA values (1.00 and 0.00) immediately distinguish a system that generates neomorphemes correctly from one that does not; System C, with reduced coverage and partial accuracy, yields the expected intermediate value of CWA $= 2/4 = 0.50$.

Mis-Generation (MIS). The final metric monitors inappropriate neomorpheme usage: cases where systems apply neomorphemes to words that should not feature them. This can occur when systems over-generalize, applying neomorphemes to words that do not refer to human entities (e.g., *tavol** instead of *tavolo*, EN: 'table') or to words that are already gender-neutral. Such mis-generations compromise translation intelligibility and represent a distinct failure mode from simply failing to generate neomorphemes where appropriate.

$$\text{MIS} = \frac{\textit{found} - \textit{correct}}{\textit{annotations}} \tag{5.4}$$

A high mis-generation rate signals problematic behavior even when accuracy appears acceptable, as it indicates that the system has not learned the appropriate scope of neomorpheme application. This metric complements the others by capturing cases where models incorrectly apply neomorphemes, rather than cases where they fail to apply them. Example D illustrates this failure mode: it achieves perfect COV and ACC by generating neomorphemes at all four annotated positions, yet also applies a neomorpheme to *dipartimento* (EN department), a non-human noun, yielding MIS $= (5-4)/4 = 0.25$.

Together, these four metrics provide a comprehensive evaluation framework for neomorpheme-based inclusive translation. Coverage and accuracy diagnose specific capabilities, coverage-weighted accuracy enables system comparison, and mis-generation flags inappropriate over-application. Unlike the general-purpose evaluation methods explored earlier in this Chapter, which assess whether translations avoid unnecessary gender marking, these metrics address the complementary challenge of evaluating whether systems can produce innovative gender-

inclusive forms when explicitly required. The generation experiments employing Neo-GATE and these metrics are presented in §6.3, where the investigation turns to whether LLMs can be prompted to produce translations with neomorphemes.

Key Points

- **Neomorpheme Metrics**: For evaluating neomorpheme-based translation, four specialized metrics are defined: coverage (proportion of expected words found, regardless of gender), accuracy (proportion with correct neomorpheme form), coverage-weighted accuracy (combined measure for system comparison), and mis-generation (inappropriate neomorpheme application). These metrics isolate neomorpheme generation capability from general translation quality.
- **Complementarity**: Coverage and accuracy separate two distinct capabilities: generating expected target vocabulary and realizing that vocabulary with appropriate gender forms (the neomorpheme generation task). Mis-generation captures a distinct failure mode where systems over-generalize by applying neomorphemes to words not referring to human entities, compromising intelligibility even when accuracy appears acceptable.

The experiments demonstrate that LLMs can serve as effective evaluators of GNT across multiple languages. LLMs achieve strong accuracy in both target-only and source-target evaluation scenarios, with GPT-4o outperforming the dedicated Italian classifier while generalizing to Spanish and German. Prompting for phrase-level annotations before sentence-level judgments consistently improves evaluation accuracy across all models and languages, indicating that prompting for intermediate analytical steps significantly improves performance on complex evaluation tasks. Crucially, the LLM-as-a-Judge approach enables source-target evaluation that can assess whether gender neutralization was appropriate given the source context, addressing a key limitation of the reference-based contrastive method and the classifier-based evaluation. Models show varying performance across languages, with German and Spanish generally yielding higher accuracy than Italian, possibly reflecting differences in training data distribution or morphological complexity.

These findings establish the LLM-as-a-Judge paradigm as a viable and scalable approach to GNT evaluation. Unlike the classifier-based method presented in §5.2, which requires synthetic data generation and language-specific fine-tuning, LLM-based evaluation can be extended to new languages through prompting alone. The experiments in §7.2.4 further validate

this approach in a real-world application context by demonstrating its flexibility in adapting to a different evaluation schema, proving its practical utility for gender-neutral rewriting in industrial settings. This scalability and flexibility, combined with the ability to incorporate source context for assessing neutralization appropriateness, makes LLM-as-a-Judge a promising framework for advancing GNT research and system development. These conclusions are grounded in the controlled, single-phenomenon settings represented in mGeNTE. How LLM-based evaluators handle more complex outputs involving multiple interacting gender cues is a natural direction for future work(§8.3.4).

Beyond the general-purpose evaluation methods explored in this Chapter, the specialized metrics for neomorpheme generation (§5.4) address the distinct evaluation requirements of innovative gender-inclusive strategies. By isolating neomorpheme generation from general translation quality, these metrics enable targeted assessment of systems' ability to produce novel inclusive forms.

Chapter 6

Generating Gender-Inclusive Translations

The previous Chapters established the resources and methods needed to evaluate gender-inclusive translation: GeNTE, mGeNTE, and Neo-GATE provide benchmarks for assessing whether systems produce appropriate outputs (§4), while the classification and LLM-as-a-Judge methods enable automated evaluation across languages (§5). With evaluation infrastructure in place, this Chapter turns to investigating the generation of gender-inclusive translations through systematic experiments, responding to **RQ4**: *How can gender-inclusive translation be automatically generated?*

The investigation proceeds from establishing baselines to progressively more sophisticated approaches. Section 6.1 first evaluates current commercial MT systems and zero-shot LLM performance on GNT, revealing their near-complete failure to produce neutral outputs without explicit guidance, and then explores few-shot prompting strategies for GNT with a commercial LLM, comparing different prompt formats and analyzing both the quantity and acceptability of generated neutralizations. Section 6.2 extends the investigation to a multilingual setting, leveraging mGeNTE to evaluate open-weight LLMs across Italian, Spanish, German, and Greek, uncovering a systematic gap between models' ability to recognize when neutrality is appropriate and their capacity to produce neutral outputs. Finally, Section 6.3 investigates the more challenging task of generating translations with neomorphemes, testing whether LLMs can produce innovative gender-inclusive forms that go beyond conservative neutralization strategies. Together, these experiments provide the first systematic empirical investigation of gender-inclusive translation generation, establishing both the potential and the limitations of current approaches. The experiments presented in this Chapter were conducted at different times and employ different models accordingly. Rather than identifying a single state-of-the-art system, the goal is to investigate the viability of NMT- and LLM-based approaches for a task that was not previously feasible, characterizing their capabilities and limitations for

gender-inclusive translation.

6.1 GNT: From Baselines to Few-Shot Prompting

Individual studies have indicated that current MT systems are ill-equipped to handle neutrality [87, 410, 259], yet systematic approaches to automating GNT remain largely unexplored. As anticipated in Chapter 2, in the current landscape LLMs offer a potential solution thanks to their adaptability to perform new tasks based on explicit instructions and examples [66] (see §2.1.3). While early evaluations showed LLMs lagging behind traditional MT systems in overall translation quality [383, 482, 516], more recent works have demonstrated that LLMs have become competitive with or superior to dedicated translation systems for high-resource language pairs [506, 242, 369, 196, 114]. Beyond raw translation quality, their versatility for controlling specific aspects in the output translation has been demonstrated for several attributes [314, 402, 165, 507].

This Section presents a systematic investigation of GNT generation, proceeding in two phases. The first phase establishes whether existing systems can produce neutral outputs when translating gender-ambiguous content without any task-specific adaptation, determining the baseline upon which dedicated approaches must improve. The second phase investigates whether few-shot prompting can elicit GNT capabilities from LLMs, experimenting with different prompt formats and analyzing both the quantity and quality of the resulting neutralizations through comprehensive manual evaluation.

6.1.1 Experimental Setup

Test Data and Manual Evaluation. The experiments are conducted on GeNTE (§4.2), using the 750 English sentences from `Set-N`, which contain gender-ambiguous human referents that should be translated into Italian in a gender-neutral way. For instance, the source *I, with **all my colleagues**, wish to thank you all for the patience* should not be translated with generic masculine formulations as in *Io, con **tutti i colleghi**$_{[M]}$, desidero ringraziare **tutti**$_{[M]}$ voi per la pazienza*, but rather with a neutral one, such as *Io, con **ogni collega**$_{[each\ colleague]}$, desidero ringraziare **tutte le persone presenti**$_{[all\ the\ people\ here]}$ per la pazienza*.

Given the novelty of GNT generation as a task, this study requires fine-grained insights into both the quantity and quality of neutral outputs. A two-layered manual evaluation protocol is therefore designed, examining both neutralization success and translation acceptability. This approach captures distinctions that automatic methods cannot: the gender-neutrality classifier (§5.2) and the LLM-as-a-Judge (§5.3) provide efficient sentence-level neutrality judgments

but cannot distinguish partial neutralizations from complete failures, nor assess the linguistic quality of neutral formulations. The manual annotations also serve to validate the classifier's reliability for GNT evaluation (Appendix F).

For each system output, evaluators first assess **neutrality** using a three-label classification: fully *neutral* (N) if all gendered expressions are successfully neutralized, fully *gendered* (G) if no neutralizations occur, or *partially neutral* (P) if some but not all of the gendered expressions present in the corresponding REF-G. are neutralized. This granularity allows distinguishing complete success from partial progress, which binary automatic evaluation conflates.

For outputs classified as neutral or partially neutral, evaluators then assess **acceptability** on a four-point scale: *acceptable* (Acc) for fluent translations that adequately represent the source meaning, *somewhat acceptable* (S-Acc) for minor issues, *somewhat unacceptable* (S-Un) for notable problems, and *unacceptable* (Un) for translations that fail in fluency or adequacy. Table 6.1 provides examples of each judgment type.

		Examples	**Neut.**	**Accep.**
A	SRC	I am **pleased** to make my contribution [...]	G	–
	REF-G	Sono ***lieto***$_{[M]}$ di potere contribuire [...]		
	OUT	Sono ***lieto***$_{[M]}$ di potere contribuire [...]		
B	SRC	[...] respect for standards lies with **the judges**.	N	Acc
	REF-G	[...] è assicurato **dai**$_{[M]}$ giudici.		
	OUT	[...] spetta ***all'autorità giudiziaria***$_{[judicial\ authority]}$.		
C	SRC	May I quote three **actors** in this field.	N	Un
	REF-G	Consentitemi di citare tre **attori**$_{[M]}$ che operano in questo campo.		
	OUT	Posso citare tre ***persone***$_{[people]}$ in questo campo.		
D	SRC	**Commissioner**, I would like to congratulate **the rapporteur**.	P	S-Acc
	REF-G	***Commissario***$_{[M]}$, vorrei congratularmi con **il**$_{[M]}$ **relatore**$_{[M]}$.		
	OUT	***Commissario***$_{[M]}$, vorrei congratularmi con ***chi ha redatto la relazione***$_{[who\ wrote\ the\ report]}$.		

Table 6.1: Output examples with the corresponding English source sentences and gendered Italian references (REF-G), along with **Neut**rality and **Accep**tability annotations. Example A shows a gendered output (*lieto*$_{[M]}$). Example B shows an acceptable neutralization using a collective noun. Example C shows an unacceptable neutralization where *actors* (in the sense of key players) becomes the overly generic *persone* (EN: people). Example D shows a partially neutral output where one term is neutralized but another remains gendered.

Three Italian native speakers, all highly familiar with neutral language practices, serve as evaluators. To enable evaluation, annotators receive the GeNTE source sentences and gendered reference translations, allowing them to identify which terms require neutralization and to judge adequacy with respect to the source meaning. Annotators focus specifically on

	BLEU	chrF	BLEURT	COMET
Amazon	<u>31.04</u>	<u>57.54</u>	<u>82.84</u>	<u>84.07</u>
DeepL	30.75	56.30	82.80	83.90
GPT-4	25.08	51.94	80.56	82.60

Table 6.2: General MT quality results for English→Italian translation with all BASELINE models. The test was performed on the Europarl common test set and computed with standard MT metrics. The best performance reported by each metric is underlined.

the portions of each sentence requiring neutralization, setting aside other aspects of translation quality. Detailed guidelines, created by the same linguist who designed the prompt examples, are provided to ensure consistency across annotations.

Systems. For the BASELINE evaluation, two popular commercial MT systems are selected: Amazon Translate[1] and DeepL.[2] For comparison with LLMs, GPT-4[3] [3] is included, which has achieved promising results in translation tasks [228], particularly for high-resource languages [383, 341]. As an instruction-following model [331, 90], GPT-4 is well-suited to adhere to provided guidance when performing tasks, which is essential for controlling gender expression in translation. The same GPT-4 model is then used in the GNT-PROMPTING setting for few-shot prompting experiments.

BASELINE Settings. Following Peng et al. [341], for GPT-4 in the BASELINE condition the simple translation prompt *"Please provide the Italian translation of the following sentence:"* is adopted, with `temperature` set to 0. Empirical studies have shown that MT is particularly sensitive to `temperature` increases, with near-zero values yielding optimal results in larger models [271] as higher values have been shown to degrade translation quality. A 0 `temperature` makes LLM generation deterministic, thus facilitating reproducibility.

Before examining GNT capabilities, the general translation quality of all systems is assessed on the Europarl [246] common test set[4] using four standard MT quality metrics: BLEU, chrF, BLEURT, and COMET. All metrics are computed with default settings. Table 6.2 reports these results, confirming that GPT-4 exhibits good cross-lingual capabilities but does not match traditional MT models in overall quality in a zero-shot prompting setting.

[1] `https://aws.amazon.com/translate/`.
[2] `https://www.deepl.com/translator`.
[3] Model `gpt-4-0613`.
[4] Retrieved from `https://www.statmt.org/europarl/`.

6.1.2 GNT-PROMPTING

To elicit GNTs in the GNT-PROMPTING setting, three prompts are designed that reflect different conceptualizations of the task. Each prompt includes a task definition emphasizing the use of gender-neutral language for human referents, followed by exemplar sentences demonstrating the expected behavior. As English has emerged as the most effective language for prompting multilingual models [421, 516, 273], English instructions are used in all prompts, with Italian appearing only in the translation examples within the demonstrations. Table 6.3 illustrates the three prompt formats.

Prompt Formats. The first template, **`Contr`** (contrastive), pairs each English source sentence with both a gendered Italian translation and a neutral Italian translation, without additional verbalized instructions. This simple format relies on the model inferring the task from the contrastive examples alone. Prior work has shown promising results with this approach for controlling binary gender expression in translation [398].

The second template, **`CoT-src`** (chain-of-thought, source-focused), structures demonstrations as question-answer pairs that break the task into intermediate reasoning steps [501, 250, 421]. This prompt first guides the identification of *source* expressions that would correspond to gendered terms in Italian, then elaborates on the neutralization of each term before providing the final translation. The reasoning process is thus anchored in the analysis of the source sentence.

The third template, **`CoT-tgt`** (chain-of-thought, target-focused), follows a similar step-by-step structure but with different intermediate steps. This prompt first provides an intermediate gendered translation, then identifies the *target* terms requiring neutralization, and finally produces the neutral translation. The reasoning process here operates on the target language, potentially making the neutralization patterns more explicit.

Task Demonstrations. Each prompt is instantiated with three exemplar sentences drawn from the institutional domain, a context where gender-neutral language is increasingly employed (see §3.1.1) and which aligns with the Europarl-based GeNTE test set. To verify whether GPT-4 can generalize from the provided examples rather than simply reiterating term-level mappings, two sets of exemplars that differ only in the gendered terms they contain are tested.

The *seen* set (`S`) includes terms that occur more than 20 times in the GeNTE test subset: *MEPs*, *President*, *everyone*, *politicians*, and *fishermen*. The *not seen* set (`NS`) includes terms that never appear in the test set: *writers*, *manager*, *employees*, *musicians*, and *freshmen*.

`Contr`	[English]: Secondly, how far does it increase transparency and accountability **of the writers**? [Italian, gendered]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**? [Italian, neutral]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?
`CoT-src`	Q: Translate the following English sentence into Italian using a gender-neutral language to refer to human entities: [Secondly, how far does it increase transparency and accountability **of the writers**]. Think step by step. A: In the English sentence there is one expression which refers to human entities and could be translated in a non-neutral way: <of the writers>. A gender-neutral translation of <of the writers> is <di chi scrive>. The final gender-neutral translation is [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?]
`CoT-tgt`	Q: Translate the following English sentence into Italian using a gender-neutral language to refer to human entities: [Secondly, how far does it increase transparency and accountability **of the writers**?]. Think step by step. A: The English sentence can be translated as [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**?]. There is one «expression with <non-neutral terms>» that refers to human entities: «<degli scrittori>». A gender-neutral alternative to «<degli scrittori>» is «di chi scrive». The final gender-neutral translation is [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?].

Table 6.3: Examples of each prompt template. The source *of the writers* is translated as *degli*$_{[M]}$ *scrittori*$_{[M]}$ in the gendered formulations and neutralized as *di chi scrive*$_{[\text{of who writes}]}$. `CoT-tgt` and `CoT-src` templates are structured as Questions and Answers. The final GNTs are highlighted.

Table 6.4 lists these term pairs with their Italian equivalents, whereas all exemplars are reported in Table D.3 in Appendix D.2. This setup is designed to investigate whether GPT-4 can generalize the neutralization principles demonstrated in the exemplars to novel gendered terms. If the model performs similarly across both conditions, this indicates that it has learned the underlying strategy rather than simply memorizing term-level mappings from the examples. Conversely, a performance drop in the `NS` condition would suggest that the model relies heavily on surface-level pattern matching, limiting its applicability to terms explicitly shown in the prompt.

The creation of exemplar sentences follows a process designed to ensure quality and consistency. Initial parallel sentences are selected from the Europarl English-Italian test set, excluding any entries already included in GeNTE. Source and reference translations are

Seen		Not seen	
en	**it**	**en**	**it**
MEPs	parlamentari europei	writers	scrittori
President	Signora Presidente	manager	direttore
everyone	tutti	employees	impiegati
politicians	politici	musicians	musicisti
fishermen	pescatori	freshmen	studenti del primo anno

Table 6.4: Source English and target Italian pairs of *seen* and *not seen* terms used in the exemplar sentences.

then modified to include the pre-selected gendered terms. For each sentence pair, GNTs are produced by a linguist experienced with Italian neutral language strategies. The resulting six exemplar sentences (three per term set) and their neutral translations are approved by all evaluators before proceeding with experiments.

Prompt format	Tokens
`Contr_S`	294
`Contr_NS`	304
`CoT-src_S`	560
`CoT-src_NS`	568
`CoT-tgt_S`	743
`CoT-tgt_NS`	781

Table 6.5: Token counts for each prompt configuration.

Table 6.5 reports the length of each prompt configuration, calculated via OpenAI's tokenizer.[5] The `Contr` prompts are most concise at approximately 300 tokens, while the chain-of-thought prompts require double the amount of tokens due to their explicit reasoning steps. Interaction with GPT-4 occurs via the chat completions API, including the complete prompt content and the input source sentence in a single message with the `user` role. To facilitate reproducibility, `temperature` is set to 0.0 throughout. Since the chain-of-thought prompts produce intermediate reasoning steps alongside the final translation, GPT-4's outputs are post-processed to extract only the final neutral translations for evaluation.

6.1.3 Results and Analysis

The analysis covers the BASELINE systems (Amazon Translate, DeepL, GPT-4 without dedicated prompting) and the six few-shot prompting configurations of GPT-4 (`Contr`, `CoT-src`,

[5] See `https://platform.openai.com/tokenizer`.

CoT-tgt prompts, each with S and NS exemplars). For each of the nine system configurations, 200 output sentences are randomly selected from the 750 outputs. The same sentences are evaluated across all conditions to enable direct comparison, yielding 1,800 annotated outputs in total. Each sentence is evaluated by one annotator, with 10% overlap across all three raters to assess inter-annotator agreement (IAA). For neutrality judgments (G, N, P), Fleiss' kappa [143] reaches 0.89, corresponding to "almost perfect agreement" [257]. All disagreements are attributable to oversights and are therefore reconciled.

For acceptability annotations, IAA is measured using the intraclass correlation coefficient (ICC) [142, 423], which accounts for the ordinal nature of the scale by capturing the distance between ratings rather than requiring exact matches. The ICC of 0.48 indicates moderate agreement, suggesting that acceptability judgments involve greater subjectivity than neutrality classification. This finding aligns with the inherent complexity of GNT anticipated in §3.3.3: multiple valid neutralization strategies exist for most sentences, and evaluators may weight the trade-offs between fluency, adequacy, and neutrality differently. To acknowledge this variability, acceptability disagreements are not reconciled.

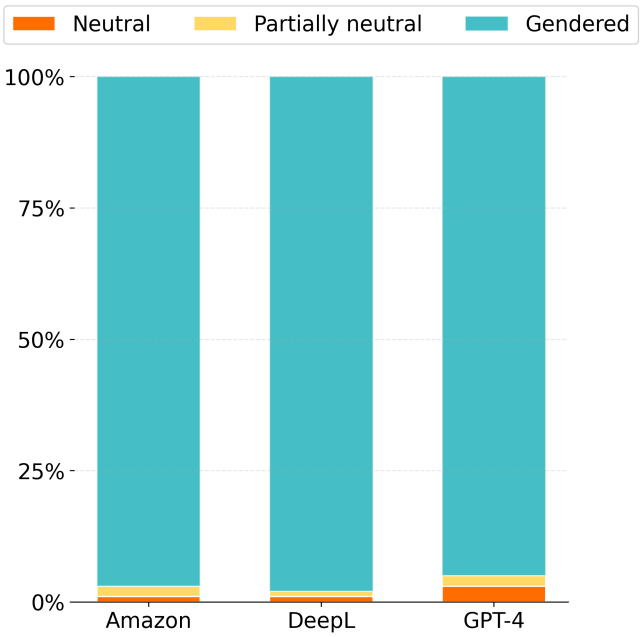

Figure 6.1: Manual evaluation of neutrality for baseline systems on 200 randomly selected GeNTE entries.

BASELINE Analysis. Figure 6.1 reports the distribution of neutrality judgments for all baseline systems. The results confirm that, used out of the box, these models are unsuitable for GNT: all three systems produce only approximately 3% neutral translations (combining

fully neutral and partially neutral outputs), with roughly 97% of outputs containing only gendered terms, predominantly masculine. This near-complete failure to produce fully neutral outputs occurs despite the source sentences containing no gender information that would justify masculine defaults, confirming previous findings on the pervasive masculine bias in MT systems (see §2.3).

Qualitative analysis reveals that the sporadic neutralizations largely correspond to incidental outcomes of literal translation choices, rather than deliberate neutralization attempts. When a system happens to generate a translation that avoids gendered expressions, it does so as a byproduct of lexical choice rather than gender-aware processing. For instance, given the source sentence *we have* ***addressed***, the Italian reference translation uses *ci siamo* ***occupati***$_{[M]}$ (literally 'we took care'), which requires masculine agreement. However, a system might instead produce ***abbiamo affrontato*** (literally 'we have faced'), which incidentally avoids gender marking because the verb *affrontare* does not require past participle agreement in this construction. The few neutralizations that do occur are judged acceptable by evaluators, but their negligible frequency and accidental nature confirm that current systems lack any systematic capability for GNT, establishing that dedicated approaches are necessary.

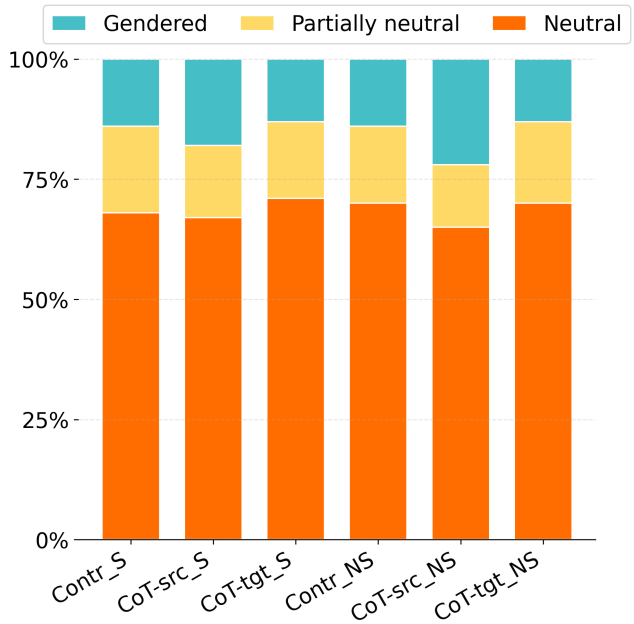

Figure 6.2: Neutrality distribution for GNT-PROMPTING configurations.

GNT-PROMPTING Neutralization Analysis. Figure 6.2 presents the distribution of neutrality judgments for all GNT-PROMPTING configurations. The results reveal a dramatic improvement over the baseline condition: across all configurations, GPT-4 produces approximately

65–70% fully neutral translations (N) and an additional 15% partially neutral translations (P), compared to the mere 3% neutral outputs observed without dedicated prompting.

Comparing prompt templates, CoT-src shows slightly lower GNT performance than the other two formats: the absence of an intermediate gendered translation in this prompt possibly had a negative impact on the model's ability to identify and transform gendered expressions. The Contr and CoT-tgt templates, which both include explicit gendered translations as reference points, achieve comparable and slightly higher neutralization rates.

Crucially, no notable differences emerge between the S (seen) and NS (not seen) exemplar conditions across any template. This finding demonstrates that GPT-4 successfully generalizes neutralization principles to newly encountered gendered terms rather than simply memorizing term-level mappings from the demonstrations. The model appears to learn the underlying task of identifying and neutralizing gendered references to humans, applying this capability to vocabulary it has not seen in the prompting examples.

GNT-PROMPTING Acceptability Analysis. Figure 6.3 presents the acceptability distribution for neutral and partially neutral outputs in the few-shot configurations. The results are generally positive: the best configurations produce over 60% acceptable neutralizations that preserve both fluency and adequate source meaning, as illustrated by example B in Table 6.1.

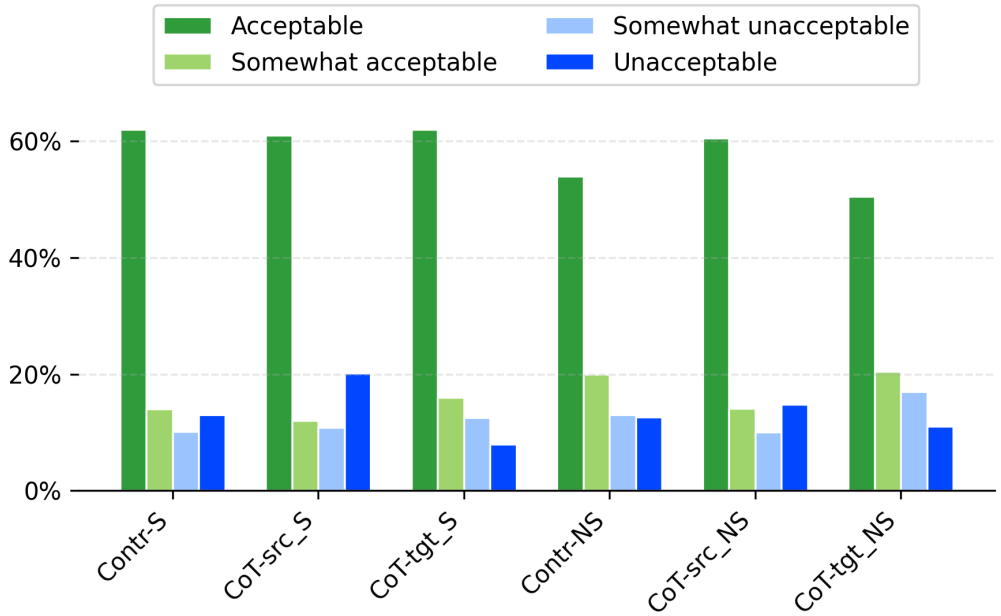

Figure 6.3: Acceptability distribution for neutral and partially neutral outputs in GNT-PROMPTING configurations.

However, a considerable proportion of outputs fall into the intermediate categories. Approx-

imately 20–30% of neutral outputs are judged as borderline (S-Acc or S-Un), reflecting cases where neutralization proves difficult to achieve without compromising fluency or adequacy. Example D in Table 6.1 illustrates this challenge: translating *rapporteur* as *chi ha redatto la relazione*$_{\text{[who wrote the report]}}$ achieves neutrality but shifts the meaning, since a rapporteur is the person responsible for reporting, not necessarily the one who wrote the report.

The use of NS exemplars appears to slightly reduce acceptability compared to S exemplars, though the effect is modest. This may reflect increased difficulty when the model must generalize further from the provided examples, occasionally producing neutralizations that are technically correct but stylistically unfit.

The Subjectivity of GNT Quality Assessment The moderate IAA [26] on acceptability (ICC = 0.48) reflects genuine complexity and subjectivity in evaluating GNT quality [44, 468, 349], a challenge that extends to evaluating gender-inclusive outputs more broadly [446]. Consider the following example:

> src: Paramilitary groups have stepped up the murders of **journalists** and human rights **activists**...
>
> out: I gruppi paramilitari hanno intensificato gli omicidi di **persone che lavorano nel giornalismo**$_{\text{[people working in journalism]}}$ e **persone attive nella difesa dei diritti umani**$_{\text{[people active in human rights defense]}}$⋯

Two evaluators judged this output as S-Acc and S-Un due to the repetition of *persone*, which they found stylistically awkward. The third evaluator rated it Un, citing adequacy concerns: *working in journalism* does not necessarily imply being a *journalist*. This example illustrates the different weights evaluators assign to the competing criteria of fluency, adequacy, and successful neutralization.

These findings have implications beyond evaluation methodology. The inherent subjectivity of GNT quality assessment suggests that future systems should account for multiple valid solutions rather than targeting a single 'correct' neutralization [352, 146, 349, 233]. Users may have different preferences regarding the trade-offs between preserving exact meaning and achieving complete neutrality, and systems might benefit from offering multiple neutralization options [339, 151, 150].

A comparison of manual annotations of neutrality with the gender-neutrality classifier (§5.2) confirms that automatic evaluation correlates well with human judgments for system-level ranking (Kendall's $\tau = 0.91$), though the three-way manual annotation captures nuances that binary classification cannot (see Appendix F for details).

The dramatic improvement from baseline to few-shot prompting demonstrates that GNT is feasible with current LLM technology when appropriate task guidance is provided. This finding calls for deeper investigation of how GNT capabilities generalize across different target languages and whether open-weight alternatives can achieve comparable performance. The next Section investigates these aspects by investigating GNT into Spanish, German, and Greek in addition to Italian, using both open-weight and commercial models.

The subjective tension between neutrality and acceptability, also reflected in the moderate inter-annotator agreement on acceptability, carries implications beyond evaluation methodology. This variability in user preferences, combined with the fact that different contexts may call for different strategies, suggests that future systems should move beyond single-output generation toward interactive approaches that allow users to specify their preferences or choose among alternative formulations (§8.4.3). These concerns should be central to future research about gender-inclusive MT evaluation and system development (§8.4.4). Additionally, the sentence-level focus of these experiments and the controlled, single-phenomenon constructions of GeNTE (§4.2.2), leaves open questions about how neutralization decisions propagate across longer and more complex texts, where consistency in referring to the same entities becomes essential (see §8.3.1 and §8.3.4).

Key Points

- **Baseline vs. prompting:** Current MT systems produce gendered outputs when translating gender-ambiguous content, with rare neutralizations occurring only incidentally. Few-shot prompting dramatically changes this: GPT-4 with dedicated prompts achieves 65–70% fully neutral translations.
- **Prompt design:** Contrastive and target-focused chain-of-thought templates outperform source-focused reasoning, suggesting that including an intermediate gendered translation helps the model identify expressions requiring neutralization.
- **Generalization capability:** GPT-4 successfully applies neutralization principles to terms not seen in the prompt exemplars, demonstrating that the model learns the underlying task rather than memorizing term-level mappings.
- **Evaluation subjectivity:** GNT quality assessment involves inherent complexity, as evaluators weigh trade-offs between fluency, adequacy, and neutralization differently. This suggests that future systems should account for multiple valid solutions rather than targeting a single correct output.

6.2 Multilingual Perspectives on GNT

The previous Section established that commercial MT systems fail almost entirely at GNT, producing gendered outputs for approximately 97% of gender-ambiguous sources. In contrast, few-shot prompting with GPT-4 demonstrates that LLMs can achieve 65–70% fully neutral translations when provided with appropriate task guidance. These findings motivate further exploration of LLM-based approaches while raising questions about generalization: the experiments thus far focus on English→Italian, a single language pair that, while representative of the challenges posed by grammatical gender languages, cannot reveal how GNT capabilities vary across different target languages. Understanding cross-linguistic variation is essential for developing MT systems that produce inclusive outputs regardless of target language, as different grammatical gender systems offer distinct neutralization resources and present unique challenges (see the cross-linguistic observations in §4.2.5).

This Section extends the investigation to a multilingual setting, leveraging mGeNTE to conduct the first systematic evaluation of GNT generation across four target languages: Italian, Spanish, German, and Greek. Using open-weight instruction-tuned LLMs, the experiments investigate two distinct capabilities: whether models can *recognize* when neutrality is appropriate based on the source sentence, and whether they can *generate* neutral translations when neutrality is called for. This distinction reflects the task definition and desiderata outlined in §3.2 and proves crucial for better understanding models' suitability for GNT, as the analysis uncovers a systematic gap between recognition and generation that carries significant implications for deploying LLMs in inclusive translation scenarios.

6.2.1 Experimental Framework

Models. The experiments employ five open-weight multilingual instruction-following models spanning different sizes and model families: Llama 3.1 8B and Llama 3.3 70B [287], Qwen 2.5 72B [360], Gemma 2 9B [454], and Phi 4 14B [2]. These models were selected from an initial pool of ten state-of-the-art multilingual LLMs spanning various model families, including Qwen, Llama, Mistral, Gemma, Phi, and Falcon [13]. Selection is based on two criteria: translation quality, assessed using xCOMET[6] [184], and format adherence, measuring whether models produce outputs conforming to the expected structured format. Five models

[6]xCOMET extends COMET by integrating error span detection capabilities alongside sentence-level scoring. While COMET provides only a single quality score per translation, xCOMET additionally identifies and categorizes translation errors according to the multidimensional quality metrics typology, labeling error spans as minor, major, or critical. xCOMET achieves state-of-the-art performance in general quality and error span detection evaluation.

are excluded from the main experiments. Falcon 3 7B, Mistral 7B, and Qwen 2.5 7B achieve comparatively lower translation quality scores, particularly on Greek; Tower Instruct 7B produces xCOMET scores below 0.4 on Greek; EuroLLM 9B largely fails to adhere to the output format requirements.

The retained models achieve high average xCOMET scores across most language pairs: 0.96 for English→German, 0.95 for English→Spanish, and 0.95 for English→Italian. Greek, as a lower-resource language with a distinct script, shows comparatively lower average performance at 0.83, reflecting the challenges of extending multilingual models to less-represented languages.

Prompt Design. The prompt structure, illustrated in Figure 6.4, comprises four components: a system prompt defining the model's role as a translator specialized in gender-neutral language, a *preamble* specifying the task rules, language-specific GNT guidelines, and four task demonstrations. The prompt instructs models to translate using gendered language when the source clearly indicates gender for human referents, and to use gender-neutral language when the source does not indicate gender. The guidelines provide concrete neutralization strategies adapted to each target language, such as using neutral synonyms, collective nouns, and neutral rephrasings while avoiding masculine generics and neomorphemes.

The four exemplar sentences are drawn from the mGeNTE parallel set (§4.2.5), with two gendered examples from `Set-G` and two neutral examples from `Set-N`. These exemplars are excluded from evaluation. Models are expected to produce structured output comprising a label indicating the source category (`GENDERED` or `NEUTRAL`) followed by the translation.

The multi-component prompt structure enables a more fine-grained analysis than the template-based comparison in §6.1.2. While the earlier experiments varied the overall reasoning structure of prompts, the present analysis systematically ablates individual prompt components to understand their contribution to GNT performance. Four configurations are tested: including both the system prompt and guidelines (`G+S`), including only the guidelines (`G`), including only the system prompt (`S`), or excluding both (`None`). This yields $4 \times 5 = 20$ model-configuration combinations per language pair.

Evaluation. Two distinct aspects of model performance are assessed. First, **source category recognition** measures the accuracy of the generated `GENDERED` / `NEUTRAL` label against the gold annotations in mGeNTE, indicating whether models correctly identify when neutrality is appropriate.

Second, **GNT accuracy** measures whether models produce correctly gendered or neutral translations as appropriate for each source sentence. For this evaluation, the LLM-as-a-

Sys.	You are a helpful `{lang}` translator specialized in gender-neutral language.
Preamble	Translate the following sentences from English into `{lang}` following these rules: 1. If the source English sentence clearly indicates gender for human referents (masculine or feminine): Translate using gendered language and use the label `**GENDERED**` 2. If the source English sentence does not indicate gender for human referents: Translate using gender-neutral language and use the label `**NEUTRAL**`
Guidelines	Guidelines for Gender-Neutral Translation: • Use neutral synonyms • Use neutral collective nouns • Use neutral rephrasings • Avoid masculine forms for generic referents • Avoid neomorphemes • Avoid double feminine/masculine forms
Exemplars	**user:** <en> `{English source}` **assist.:** <`{lang}`> `**GENDERED**` `[{gendered translation}]` **user:** <en> `{English source}` **assist.:** <`{lang}`> `**GENDERED**` `[{gendered translation}]` **user:** <en> `{English source}` **assist.:** <`{lang}`> `**NEUTRAL**` `[{neutral translation}]` **user:** <en> `{English source}` **assist.:** <`{lang}`> `**NEUTRAL**` `[{neutral translation}]`

Figure 6.4: GNT prompt overview with labeled sections. The prompt consists of **Sys**tem instructions, a **Preamble** with translation rules, **Guidelines** to achieve gender-neutrality, and **Exemplars** provided as conversational turns.

Judge framework presented in §5.3 is employed. The optimal prompt configuration from that framework (CROSS-P+L), which elicits phrase-level annotations before sentence-level judgments and achieved the best results in the Italian, Spanish, and German experiments, is adapted to all four languages covered in mGeNTE, including Greek. To validate this evaluation setup on LLM-generated translations (as opposed to the human-written text used in Chapter 5), 1,000 model outputs are manually annotated and two LLM evaluators are tested: Qwen 2.5 72B Instruct and GPT-4o. Table 6.6 reports the validation results. GPT-4o emerges as the

best-performing evaluator, achieving 0.92 accuracy and 0.87 macro F1 overall, with Spanish yielding the highest per-language performance (0.96 accuracy, 0.94 macro F1) and Greek the lowest (0.89 accuracy, 0.80 macro F1). Based on these results, GPT-4o with the CROSS-P+L prompt is adopted for all GNT accuracy evaluations in this Section.

Model	Language	Accuracy	Macro-F1
Qwen 2.5 72B	German	0.84	0.79
	Greek	0.88	0.78
	Italian	0.87	0.78
	Spanish	0.90	0.86
	average	0.85	0.80
GPT-4o	German	0.89	0.85
	Greek	0.89	0.80
	Italian	0.92	0.87
	Spanish	0.96	0.94
	average	**0.92**	**0.87**

Table 6.6: Validation results for the LLM-as-a-Judge GNT evaluation on 1,000 manually annotated model outputs, reported overall and per language pair.

6.2.2 Results of Source Gender Recognition

Figure 6.5 (left panels) presents source category recognition results across all models and language pairs. The findings reveal strong, consistent performance: models reliably distinguish gender-ambiguous source sentences (`Set-N`) from those containing explicit gender cues (`Set-G`).

Label accuracy remains high across languages, models, and prompt configurations, with only minor variance. Performance on `Set-G` (identifying gendered sources) is slightly higher than on `Set-N` (identifying ambiguous sources), but both exceed 85% for most model-language combinations. Even the smallest model (Llama 3.1 8B) achieves respectable recognition accuracy, and larger models show only modest improvements in this dimension.

This consistent performance suggests that recognizing when neutrality is appropriate represents a relatively tractable subtask for instruction-following LLMs. The models appear to successfully identify linguistic cues in the source sentence that indicate whether gender information is present or absent, a capability that transfers well across the typologically diverse target languages in the evaluation. This finding should be interpreted in light of the controlled, single-phenomenon sentence constructions used in mGeNTE (§4.2.2): real-world source

sentences may present more complex configurations with multiple interacting gender cues, for which recognition performance may differ (see §8.3.4).

6.2.3 GNT Results

While source category recognition proves robust, the picture changes substantially when examining actual translation outputs. Figure 6.5 (right panels) presents GNT accuracy results, revealing a systematic gap between recognizing when neutrality is needed and successfully producing neutral translations.

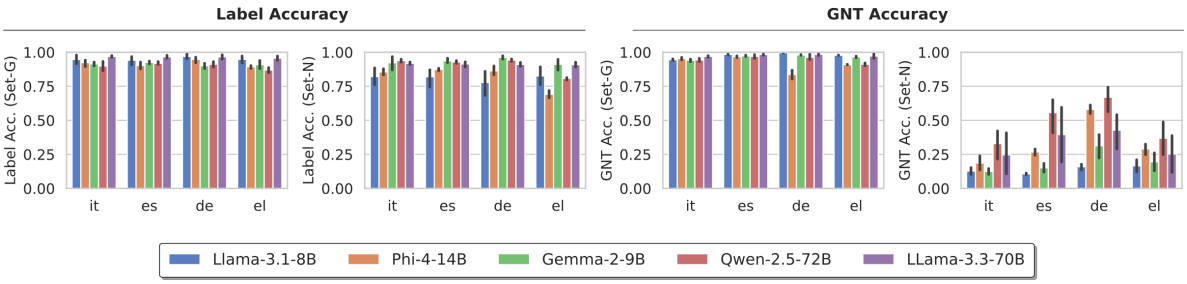

Figure 6.5: Source category (*left*) and GNT accuracy (*right*) results across mGeNTE Sets (averaged across prompt variations).

The Recognition-Generation Gap. For `Set-G` sentences, where gendered translation is appropriate, models consistently produce correct outputs with high accuracy across all languages. However, for `Set-N` sentences requiring neutral translation, accuracy drops substantially and exhibits higher variance across configurations. This asymmetry indicates that models have learned to produce gendered translations reliably but struggle to operationalize neutralization strategies at generation time. This pattern echoes findings from studies on pronoun fidelity, where LLMs demonstrate inconsistent behavior between recognizing correct pronoun usage and generating appropriate forms [167].

To quantify this gap, label-translation coherence is measured as the agreement between the generated label and the actual gender expression in the translation. As shown in Figure 6.6, models systematically produce gendered translations when assigning a `GENDERED` label, achieving near-perfect coherence. However, coherence drops sharply for `NEUTRAL` labels, often falling below random chance. This finding confirms that correct source categorization does not guarantee correct GNT: models may correctly identify that neutrality is needed yet still produce gendered outputs. Such a dissociation is expected given the asymmetry between the two tasks, recognition being a binary sentence-level classification and generation requiring targeted structural modifications. Nonetheless, the label-translation coherence results provide

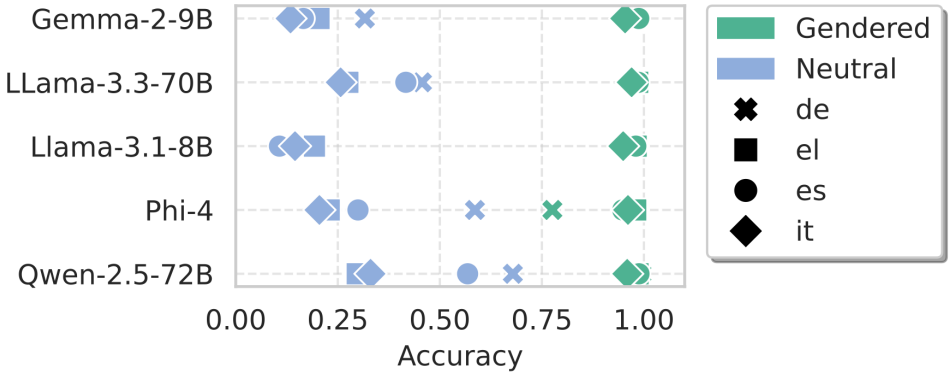

Figure 6.6: Accuracy of label-translation coherence. Reports the agreement of gender expression in the translation (gendered/neutral) with the generated label. Scores are averaged across prompt configurations.

systematic empirical evidence for this gap across typologically diverse language pairs, showing precisely how and to what degree it manifests in multilingual GNT.

Cross-Linguistic Variation. GNT performance varies substantially across target languages. English→German and English→Spanish generally achieve higher neutral translation accuracy than English→Italian and English→Greek. Greek's lower performance aligns with its reduced overall translation quality as a lower-resource language. However, the underperformance on Italian is more surprising given the models' generally solid translation quality, but mirrors the pattern observed in the multilingual LLM-as-a-Judge experiments (§5.3.2).

This cross-linguistic variation may reflect multiple factors: differences in how these languages are represented in model training data [230, 368, 278], varying availability of gender-neutral linguistic resources in each language, and potentially sociolinguistic factors such as the relative prominence of inclusive language practices in different linguistic communities. German, for instance, has seen substantial public discourse around gender-inclusive language [344, 258, 488, 160, 414], which may be reflected in training data composition.

Model Size and Prompt Sensitivity. Larger models generally outperform smaller ones on GNT, with Qwen 2.5 72B leading overall, followed by Llama 3.3 70B and Phi 4 14B. However, model size alone does not determine performance: Phi 4 14B outperforms the much larger Llama 3.3 70B on some language pairs, suggesting that architecture and training composition also play significant roles.

Prompt configuration substantially affects GNT accuracy, particularly for larger models. Including both the system prompt and guidelines (`G+S`) yields the highest accuracy, while removing both (`None`) leads to the worst performance. Notably, larger models show greater

gains from richer prompts, better leveraging the provided context to improve neutral generation. Llama 3.3 70B, for instance, drops to small-model performance levels in the `None` configuration but achieves strong results with full context. This sensitivity to prompt design has practical implications: deploying LLMs for inclusive translation requires careful attention to prompt engineering, and minimal prompts may fail to elicit the desired behavior even from capable models.

Broader Experimental Context. These experiments are part of a broader investigation that pairs behavioral evaluation with explainability analysis to shed light on the internal mechanisms underlying LLM-based GNT. The explainability component, while not central to this thesis, offers insights that help explain the observed patterns. Using feature attribution methods [140, 4, 514] to measure the contribution of each prompt component to generated outputs, the analysis reveals that source category recognition and translation generation rely on different context signals. For label generation, models primarily attend to the source sentence and the label tokens from exemplars, learning to identify gender-related cues directly from the input. For translation generation, however, models draw more heavily on the target-language portions of the exemplars and the guidelines, indicating that producing neutral formulations requires access to target-language neutralization patterns. This differential reliance on context components helps explain the recognition-generation gap discussed above: models can identify when neutrality is needed by attending to source cues, but successfully generating neutral translations requires additional target-language knowledge that may not be equally accessible across all languages and models. The finding also validates the prompt design choices showing that richer target-language context improves GNT performance. Moreover, it suggests that future work on improving LLM-based GNT should focus specifically on enhancing models' access to and utilization of target-language neutralization strategies, whether through improved training (§8.4.6) or interactive approaches that allow users to guide the generation process (§8.4.3).

The findings presented in this Section highlight a fundamental challenge for LLM-based GNT: the dissociation between understanding and execution. Models reliably detect when gender-neutrality is appropriate, demonstrating comprehension of the underlying linguistic distinctions articulated in the task definition and desiderata (§3.2). Yet this understanding does not automatically translate into appropriate generation behavior. The recognition-generation gap suggests that producing neutral formulations requires capabilities beyond task recognition, potentially including access to neutralization strategies in the target language and the ability to apply them while maintaining translation quality. This asymmetry echoes findings from the

monolingual gender-neutral rewriting experiments discussed in §7.2.4, where classification proves substantially easier than generation, while rewriting required increasingly capable models to achieve acceptable quality.

Cross-linguistic variation in GNT performance reveals that the challenges extend beyond model capabilities to language-specific factors. German and Spanish yield higher neutral translation accuracy than Italian and Greek, a pattern that cannot be fully explained by general translation quality alone, given Italian's solid baseline performance. Multiple factors may contribute: differences in how languages are represented in training data [230, 278, 80, 273], varying availability of gender-neutral linguistic resources, and sociolinguistic factors such as the relative prominence of inclusive language practices in different linguistic communities. German, for instance, has seen substantial public discourse around gender-inclusive language, with documented increases in usage across media outlets and ongoing political debates that have made it a prominent topic in public communication [487, 160, 344], which may be reflected in training data composition. These cross-linguistic differences underscore that GNT solutions may not transfer straightforwardly across languages, reinforcing the need for multilingual benchmarks (like mGeNTE) and scalable evaluation methods (like the LLM-as-a-Judge approach) that enable systematic investigation of such variations, and highlighting the importance of expanding language coverage in future work (§8.4.2).

The strong dependence on prompt design carries practical implications for deploying LLMs in inclusive translation scenarios. Larger models benefit substantially more from richer prompt contexts: including both explicit guidelines and system prompts yields the highest GNT accuracy, while minimal prompts produce poor results even for capable models. This sensitivity suggests that careful prompt engineering can partially compensate for model limitations but also indicates that robust multilingual GNT will require improvements in base model capabilities. From a deployment perspective, the results indicate that current open-weight LLMs are not yet completely reliable for automatic GNT, particularly for languages where neutral accuracy remains low, though they may serve as useful assistive tools when combined with human oversight. This calls for continued research into improved training methods and dedicated data curation strategies for gender-inclusive generation (see §8.4.6).

Key Points

- **Recognition vs. Generation Gap**: LLMs reliably identify when gender-neutrality is appropriate (source category recognition) but struggle to consistently produce neutral translations, revealing a dissociation between understanding and execution.
- **Label-Translation Coherence**: Models produce gendered translations consistently

when they identify gendered sources, but coherence drops sharply for neutral cases, confirming that correct categorization does not guarantee correct generation.

- **Cross-Linguistic Variation**: GNT accuracy varies across target languages, with German and Spanish yielding better results than Italian and Greek. This variation reflects differences in resource availability, training data composition, and potentially sociolinguistic factors.
- **Prompt Sensitivity**: Larger models benefit more from richer prompt contexts. Including explicit guidelines and system prompts substantially improves GNT accuracy, with minimal prompts yielding poor results even for capable models.

6.3 LLM Experiments with Neomorphemes

The previous Sections investigated gender-inclusive translation through a conservative approach that draws on existing linguistic resources. This Section shifts focus to the generation of translations employing gender-inclusive neomorphemes, the innovative morphological markers designed for explicit non-binary gender expression (§3.1.3). Where conservative neutralization tends to circumvent gender marking entirely, neomorpheme-based translation requires systems to produce novel word forms that are absent or extremely rare in their training data.

This requirement poses a significant challenge for neural approaches: generating correct neomorpheme forms demands not only recognition of contexts requiring gender-inclusive marking but also the ability to produce morphological patterns that fall outside the distribution of standard training corpora. The experiments presented in this Section investigate whether LLMs (both closed and open-weight) can acquire this capability through in-context learning, applying neomorphemes appropriately when provided with task demonstrations despite minimal prior exposure to these forms. The evaluation leverages the English→Italian Neo-GATE (§4.3) benchmark,[7] which was designed specifically for neomorpheme-based translation, testing multiple LLMs across different prompting configurations.

[7]The return to a single language pair reflects both practical and linguistic considerations: Neo-GATE covers only English→Italian, and neomorpheme conventions vary substantially across grammatical gender languages, with different linguistic communities adopting distinct paradigms (e.g., the Schwa in Italian, the -e ending in Spanish, the Gender Star * in German), making cross-linguistic comparison methodologically complex.

	BLEU	CHRF	TER ↓	BERTSC.	COMET
OPUS-MT	27.53	57.61	58.95	87.42	82.68
GPT-4	**32.34**	**61.11**	**54.87**	**88.76**	**87.05**
Tower	30.88	59.41	56.96	88.17	86.21
Mixtral	29.63	58.68	59.35	87.81	86.11
Llama 2	26.28	55.92	61.98	87.02	84.23

Table 6.7: Translation quality on FLORES-101 for English→Italian. Cases where LLMs outperform the MT baseline are underlined in the original evaluation.

6.3.1 Experimental Settings

The experiments described in this Section evaluate the ability of different LLMs to generate Italian translations containing neomorphemes through prompting. Four models representing different architectures and training approaches are tested on Neo-GATE, using two prominent Italian neomorpheme paradigms. The following paragraphs detail the model selection, paradigm specifications, and baseline translation quality verification.

Models. The experiments employ four LLMs representing different model families and typologies: GPT-4 (`gpt-4-0125-preview`) [3] as a commercial reference, and three open-weight models including Mixtral-8x7B-Instruct [226], Llama 2-70B-chat [462], and Tower-7B [15], a LLama-based model specifically fine-tuned for MT. All models use `temperature` 0.0 for reproducibility. Neural MT models are not included as no existing model supports neomorphemes and no dedicated training or fine-tuning data is available.

To verify the suitability of these models for translation tasks, their general English→Italian performance is first evaluated on FLORES-101 [179], by prompting the models to translate with a few-shot prompt (see Appendix D.3). Table 6.7 reports these results alongside opus-mt-en-it,[8] a state-of-the-art encoder-decoder NMT model (see §2.1.2), as reference. For this general MT evaluation, the metrics employed include BLEU, chrF, and TER for surface similarity to human-made reference translations, BERTScore for semantic adherence to those references, and COMET for semantic adherence to the source. The LLMs perform competitively, with GPT-4 and Tower often outperforming the dedicated MT system. It is worth noting that the BASELINE evaluation in §6.1.1 showed GPT-4 underperforming traditional MT systems in zero-shot settings. However, with appropriate few-shot prompting, it surpassed opus-mt-en-it, confirming recent findings that position LLMs as the new dominant paradigm for MT when provided with appropriate task guidance and context [164, 8, 506].

[8] `https://huggingface.co/Helsinki-NLP/opus-mt-en-it`

This confirms that any limitations observed in neomorpheme generation cannot be attributed to poor general translation capability.

Neomorpheme Paradigms. These experiments focus on the two most widely used Italian neomorpheme paradigms [94, 447, 144]: the **Asterisk** paradigm, which uses the symbol '*' as a graphemic substitute for gendered morphemes (e.g., *ragazz** instead of *ragazzi*$_{[M]}$/*ragazze*$_{[F]}$), and the **Schwa** paradigm, which distinguishes singular and plural forms using 'ə' and 'ɜ' respectively (e.g., *ragazzə* for singular, *ragazzɜ* for plural). The Schwa paradigm presents an additional challenge due to this number distinction, requiring models to track grammatical number alongside the inclusive gender marking.

For each paradigm a tagset mapping is created (see Appendix C.1), associating the tags used in the tagged references (see §4.3.2) with the desired form for that specific paradigm. As no complete codification of the use and the orthography of neomorphemes in Italian is available [457], established resources such as the website *Italiano Inclusivo*[9] and examples found in scientific literature, such as Rosola et al. [389], serve as references. Since these sources do not cover the whole set of possibly gendered elements in the grammar, missing forms are derived by analogy from elements of the same class. For example, since none of these sources describes the full set of articulated prepositions, which express gender in Italian, the given examples serve as a model for the rest of the class.

6.3.2 Prompting Configurations

The experiments employ zero-shot and few-shot prompting configurations to assess both models' baseline capabilities and their ability to learn from task demonstrations.

Zero-Shot. Similar to §6.1.1, the zero-shot setting, models receive only a verbalized task description instructing them to translate using neomorphemes as substitutes for gendered morphemes when referring to human entities. No examples are provided, testing whether models have any inherent capability for this task based solely on instruction comprehension and whether few-shot prompting can improve upon it [264].

Few-Shot Formats. The few-shot experiments include three prompt formats of increasing complexity, following the structure used in prior work on controlling gender expression in translation [398]. Specifically, they are based on the three prompt formats illustrated in Table 6.8, namely:

[9] See `https://italianoinclusivo.it/scrittura/`.

- The **Direct** format, which presents source sentences paired directly with their neomorpheme translations. This minimal format tests whether models can infer the task from input-output pairs alone.

- The **Binary** format, which includes both a gendered Italian translation and the corresponding neomorpheme translation for each source sentence. This format frames the task as a double-output translation, asking models to first produce a gendered translation and then a second version with neomorphemes that should be identical except for the words expressing gender. By showing the contrast between gendered and neomorpheme forms explicitly, this format makes the relationship between the two more apparent. This approach mirrors the contrastive prompting strategy that proved effective for GNT in §6.1.

- The **Ternary** format, which extends Binary by providing masculine, feminine, and neomorpheme translations for each source. The rationale for this triple-output structure is that by instantiating a ternary opposition, models may better identify which parts of the target sentence should remain identical across all three translations and which parts should differ. Framing the task this way could help models infer that the gender expressed in the third translation should be something other than masculine or feminine, potentially clarifying the role of neomorphemes as a distinct category.

All the four models used for these experiments expect prompts in a chat format, with `user` messages providing input and `assistant` messages representing the model's desired output.[10] For the few-shot prompts the experiments adhere to this structure, whereas for the zero-shot prompts only a single `user` message is provided.

Task Demonstrations. In the few-shot settings, the experiments include 1, 4, and 8 task demonstrations. These values are chosen as the minimum necessary to elicit in-context learning (1) and a compromise between a high number of demonstrations and the computational cost of inference (8).

Exemplar sentences are enclosed in angle brackets, and models are expected to reproduce this structure, facilitating automatic extraction of the final translation during post-processing.

The exemplar sentences are selected from Neo-GATE's development set. The selection targets sentences that represent the average tag *density* of the development set (i.e., the number of neomorpheme tags per reference) and that offer a balanced mix of singular and plural tags.

[10] See `https://huggingface.co/docs/transformers/main/en/chat_templating`.

This balance is particularly important for the Schwa paradigm, which distinguishes singular (ə) and plural (ɜ) forms. The prompts are then formatted using each paradigm's tagset mapping.

PROMPT	ROLE	INSTRUCTION / EXAMPLE
Zero-shot	`user`	Translate the following English sentence into Italian using the neomorpheme '*'. To do so, the neomorpheme '*' should be used as a substitute for masculine and feminine morphemes in words that refer to human beings. [English] <`{input sentence}`> [Italian]
Direct	`user`	[English] <I never buy flowers for my friends.> [Italian]
	`assistant`	<Non compro mai fiori per l* mi* amic*.>
Binary	`user`	[English] <I never buy flowers for my friends.> [Italian, gendered]
	`assistant`	<Non compro mai fiori per i miei amici.> [Italian, neomorpheme] <Non compro mai fiori per l* mi* amic*.>
Ternary	`user`	[English] <I never buy flowers for my friends.> [Italian, masculine]
	`assistant`	<Non compro mai fiori per i miei amici.> [Italian, feminine] <Non compro mai fiori per le mie amiche.> [Italian, neomorpheme] <Non compro mai fiori per l* mi* amic*.>

Table 6.8: Examples of all the prompts used in the experiments. The few-shots prompt examples include the Asterisk neomorpheme. Words expressing gender are highlighted.

6.3.3 Results and Analysis

The results for both zero-shot and few-shot experiments are presented below, analyzing model performance through the four metrics introduced in §5.4: coverage (COV), accuracy (ACC), coverage-weighted accuracy (CWA), and mis-generation (MIS).

For the few-shot experiments, each metric is reported separately to provide a detailed picture of model behavior. Notably, Llama 2 results are incomplete: in several configurations (all 8-shot settings), the model fails to reproduce the expected output format, omitting angle brackets or labels and producing outputs with long hallucinations [225, 216] that cannot be automatically post-processed. Since manual investigation of the model outputs did not reveal hints of improved performance in these settings, only the Asterisk Direct 8-shot configuration is reported for Llama 2.

	ASTERISK				SCHWA			
	COV ↑	**ACC ↑**	**CWA ↑**	**MIS ↓**	**COV ↑**	**ACC ↑**	**CWA ↑**	**MIS ↓**
GPT-4	57.08	**74.63**	**42.60**	45.78	46.91	**60.19**	**28.24**	72.77
Tower	**77.57**	0.00	0.00	**0.00**	**77.25**	0.00	0.00	**0.00**
Mixtral	35.22	37.92	13.35	52.20	30.05	27.79	8.35	61.44
Llama 2	56.72	0.57	0.32	16.70	57.60	0.35	0.20	12.79

Table 6.9: Zero-shot setting results, reporting the coverage (COV), accuracy (ACC), coverage-weighted accuracy (CWA), and mis-generation (MIS) scores.

Zero-Shot Results Analysis. Table 6.9 presents the zero-shot results, revealing markedly different behaviors across models.

GPT-4 and Mixtral achieve substantially higher accuracy than Llama 2 and Tower, with GPT-4 approximately doubling Mixtral's performance. The accuracy scores indicate that, among the terms that matched the reference vocabulary, GPT-4 correctly generates 74.63% of Asterisk neomorphemes and 60.19% of Schwa neomorphemes, while Mixtral reaches 37.92% and 27.79% respectively. When accounting for coverage, the gap widens further: GPT-4's coverage-weighted accuracy is more than three times that of Mixtral (42.60 and 28.24 versus 13.35 and 8.35).

However, both models produce considerable mis-generations, with mis-generation rates often exceeding their respective coverage scores. This finding indicates that GPT-4 and Mixtral generate numerous neomorphemes, but apply them incorrectly in the majority of cases. Regardless of these errors, both models perform better with the Asterisk paradigm than with the Schwa, possibly because the Schwa's use of distinct singular and plural forms adds complexity to an already challenging task.

In contrast, Llama 2 and Tower severely under-generate neomorphemes in the zero-shot setting. Llama 2's near-zero accuracy scores (0.57 and 0.35) combined with its low mis-generation rates (16.70 and 12.79) indicate that the model rarely generates neomorphemes, and when it does, it applies them inaccurately. Tower presents a different pattern: its high coverage scores (77.57 and 77.25) combined with zero scores on all other metrics indicate that the model produces fluent, fully gendered outputs without ever generating neomorphemes.

Tower's behavior may be attributable to the composition of its fine-tuning dataset,[11] in which neomorpheme characters are practically absent (only 3 occurrences of 'ə' in English segments, with no occurrences of 'ɜ' or '*'). This absence likely prevents the model from producing these characters even when explicitly instructed to do so [79, 368, 67]. Unfortunately,

[11] See TowerBlocks [15] at `https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.1/`.

since the training data for the other models is not publicly available, this hypothesis cannot be further investigated and no definitive conclusions can be drawn about the relationship between training data composition and neomorpheme generation capability. This limitation highlights a broader issue with 'open' models: while Tower, Mixtral, and Llama 2 make their weights publicly available, they do not disclose their training data, placing them in the 'open-weight' rather than truly 'open-source' category under *openness* frameworks [503, 289, 277].

Few-Shot Results: Coverage and Accuracy. Figure 6.7 presents coverage and accuracy scores across all few-shot configurations. Examining coverage (Figure 6.7a), **few-shot prompting generally improves coverage compared to zero-shot results**. For Mixtral and Llama 2, coverage increases with more demonstrations. Regarding prompt formats, the Direct format generally produces higher coverage, with only GPT-4 performing better with the Ternary format. The neomorpheme paradigm also affects coverage: scores are generally higher with the Asterisk paradigm than with the Schwa. As discussed below in the analysis of mis-generation, this difference can be attributed to models' tendency to produce more mis-generations with the Schwa paradigm, in part due to the increased complexity of this paradigm.

		Asterisk			Schwa		
		Direct	Binary	Ternary	Direct	Binary	Ternary
GPT-4	1	64.26	71.24	80.11	64.26	64.70	71.00
	4	69.34	68.46	74.26	72.00	68.54	72.45
	8	71.68	69.46	74.43	73.17	70.88	71.28
Tower	1	76.76	76.24	73.62	77.65	73.17	67.16
	4	78.30	75.23	70.79	76.89	69.30	64.62
	8	77.85	76.48	73.42	76.28	69.30	63.09
Mixtral	1	54.22	52.60	52.32	45.06	20.61	23.88
	4	67.37	61.64	56.03	64.54	50.58	42.23
	8	72.13	64.86	50.79	70.27	58.17	53.93
LLama 2	1	54.34	54.26	54.58	62.04	47.08	44.70
	4	62.44	61.84	59.66	66.84	57.97	52.76
	8	64.02					

(a) **Coverage** percentage scores.

		Asterisk			Schwa		
		Direct	Binary	Ternary	Direct	Binary	Ternary
GPT-4	1	70.62	30.18	4.48	50.53	47.01	24.20
	4	68.70	80.67	44.27	57.59	81.93	57.24
	8	67.02	80.49	40.11	58.82	82.93	60.50
Tower	1	1.89	2.96	6.30	1.35	3.47	13.03
	4	1.85	5.42	10.48	2.52	10.65	20.79
	8	2.80	4.85	13.24	3.65	11.06	19.25
Mixtral	1	56.18	85.66	90.21	39.66	71.23	80.24
	4	39.04	78.66	87.40	30.25	70.89	78.41
	8	28.08	73.69	82.02	17.51	62.27	72.03
LLama 2	1	6.90	6.43	6.43	4.10	8.91	7.40
	4	1.81	8.74	4.41	1.93	7.03	4.43
	8	1.95					

(b) **Accuracy** percentage scores.

Figure 6.7: Coverage and accuracy results in the few-shot settings. Darker shades indicate better performance.

Coverage alone, however, only indicates the proportion of annotated terms that models generate; it does not reveal how many of those terms include correctly formed neomorphemes. Turning to accuracy (Figure 6.7b), all models improve their performance in at least one few-shot setting compared to zero-shot, confirming the benefits of in-context learning for generative

tasks involving neologistic expressions [209, 294]. Mixtral and GPT-4 produce the highest rates of correct neomorphemes, with Mixtral reaching 90.21% accuracy and GPT-4 reaching 82.93% in their best configurations. Tower and Llama 2, despite showing improvements, remain unsuitable for the task given their persistently low scores.

Surprisingly, more demonstrations do not necessarily lead to higher accuracy. While coverage generally increases with additional examples, this positive trend holds for accuracy only with GPT-4 and Tower, indicating that these models generate more neomorphemes and do so more correctly as they receive more examples. Conversely, accuracy for Llama 2 and Mixtral *decreases* significantly with more demonstrations. Combined with their rising coverage, this pattern indicates that these models produce fewer neomorphemes and more gendered terms as demonstration count increases.

This counterintuitive behavior can be explained by considering how models learn task boundaries from demonstrations [264, 526]. The higher accuracy and lower coverage observed in the 1-shot settings for Llama 2 and Mixtral may result from fortuitous correct generations in a context of over-generation, where the models apply neomorphemes liberally (including incorrectly). As models receive more examples and better understand the task structure, they become more conservative, producing more gendered terms (higher coverage) but potentially missing opportunities for correct neomorpheme use (lower accuracy). This hypothesis is examined further through the mis-generation analysis below.

Regarding neomorpheme paradigms, Mixtral performs better with the Asterisk while Tower performs better with the Schwa, consistent with the zero-shot results. GPT-4 and Llama 2 show no consistent paradigm preference. The ability to generate one neomorpheme type more accurately than another likely depends on models' robustness to novel grammatical paradigms and on how the specific characters are represented in each model's training data [79, 368]. Unfortunately, as discussed above, this aspect cannot be investigated further as training data is not publicly available for these models, with the exception of Tower's fine-tuning dataset.

Few-Shot Results: Coverage-Weighted Accuracy. To enable direct comparison of overall model performance, Figure 6.8 presents coverage-weighted accuracy scores. This metric offers a comprehensive view of model behavior in each setting, allowing for fair comparison across different systems by accounting for both coverage and accuracy. The results confirm that all models improve their performance through few-shot prompting. The benefits of in-context learning are substantial, and there appears to be room for further improvement with additional demonstrations. GPT-4 and Mixtral emerge as the best-performing models, with the gap between them narrowing considerably compared to the zero-shot experiments. In their best configurations, GPT-4 achieves 58.78 CWA (Schwa, Binary, 8 shots) and Mixtral achieves

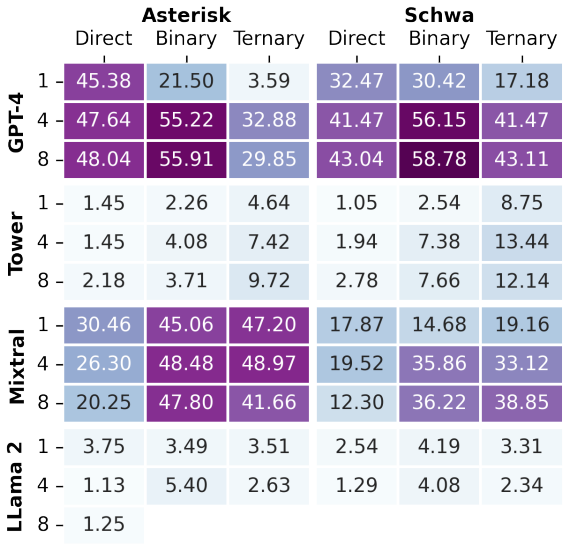

Figure 6.8: Coverage-weighted accuracy percentage scores for the few-shot settings. Darker shades indicate better performance.

48.97 CWA (Asterisk, Ternary, 4 shots).

Examining prompt format effects, GPT-4 generally performs best with the Binary format and 4 or 8 shots, while Mixtral achieves optimal results in the Binary/Ternary region with 4 or 8 shots, particularly with the Asterisk paradigm. These patterns suggest that the contrastive structure provided by Binary and Ternary formats helps in modelling the relationship between gendered and neomorpheme forms, echoing the benefits of contrastive prompting observed in the GNT experiments (§6.2.3).

Few-Shot Results: Mis-Generation analysis. The preceding analysis focused on correct neomorpheme generation for annotated human referents. To gain a more complete picture of model behavior, it is necessary to consider mis-generation: the inappropriate application of neomorphemes to terms that should not receive them. The results are presented in Figure 6.9.

Mixtral produces the most mis-generations, particularly in the 1-shot and 4-shot Binary and Ternary configurations. Table 6.10 illustrates typical mis-generation patterns from Mixtral's outputs. In the first example, Mixtral applies the Schwa to the verb *spero* ('I hope'), which does not refer to a human entity and should not be neutralized. In the second example, the model applies neomorphemes to the preposition *a*, the verb *rimanere* ('to remain'), and the noun *silenzio* ('silence'), none of which require gender-inclusive marking.

These examples confirm the hypothesis presented in the coverage and accuracy analysis regarding the interaction between these two metrics. In low-shot settings, Mixtral over-generates neomorphemes, applying them indiscriminately to many words. This behavior

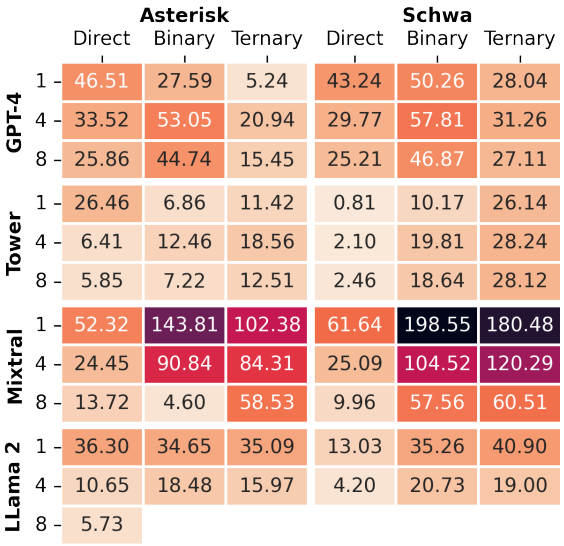

Figure 6.9: Mis-generation percentage scores for the few-shot settings. Higher scores (darker shades) indicate worse performance.

produces both correct generations (when neomorphemes happen to be applied to appropriate terms) and mis-generations (when applied to inappropriate terms). The high accuracy paired with low coverage in these settings reflects this pattern: by over-generating neomorphemes, Mixtral produces fewer standard gendered words (which contribute to coverage) and many neomorpheme-containing words that are either correct or mis-generations.

With more demonstrations, Mixtral generates significantly fewer mis-generations, and while its accuracy decreases, its coverage improves. This indicates that additional examples help the model learn the appropriate scope of neomorpheme application, producing better-formed outputs overall even if the rate of correct neomorpheme use among generated terms decreases. Mixtral's behavior thus demonstrates how the mis-generation metric complements the analysis of model performance, shedding light on unwanted phenomena that coverage and accuracy alone cannot signal.

Llama 2 exhibits a similar pattern to Mixtral, producing more mis-generations with fewer demonstrations and progressively improving with more examples. GPT-4 and Tower show the opposite trend: they generate fewer mis-generations overall, with Tower producing even fewer than GPT-4. However, the best-performing configurations for both models are also those with the highest mis-generation rates. This suggests a trade-off between attempting neomorpheme generation (which increases both correct generations and mis-generations) and conservative behavior (which minimizes errors but also limits correct outputs). Future work should focus on improving the ratio of correctly generated neomorphemes to total neomorphemes, enabling models to be both productive and precise.

Source	I hope the shaman can help us.
Annotation	lo la lə; sciamano sciamana sciamanə;
Output	**Sperə** che lə sciamanə possa aiutarci.
Source	They asked everyone to remain silent.
Annotation	tutti tutte tuttɜ;
Output	Hanno chiesto a **tɜ** di **rimanerə** in **silenziə**.

Table 6.10: Examples of mis-generation found in Mixtral's Schwa, 1 shot, Binary prompt outputs. Words containing neomorphemes are underlined, mis-generations are in bold.

These experiments establish the first systematic evaluation of LLM capabilities for neomorpheme-based gender-inclusive translation, revealing both promising potential and significant challenges that distinguish this task from conservative neutralization.

Among these challenges, the choice of neomorpheme paradigm affects task difficulty. The Schwa paradigm introduces additional challenges compared to the Asterisk, as reflected in its generally higher mis-generation rates. While the Asterisk uses a single character for all contexts, the Schwa requires tracking grammatical number alongside gender-inclusive marking, distinguishing singular (ə) from plural (ɜ) forms. This added complexity increases the likelihood of incorrect application, particularly for models still learning the morphological constraints of neomorpheme use. From a practical standpoint, simpler paradigms may prove more amenable to LLM-based generation, though the ultimate choice should be guided by community preferences (§8.4.5), accessibility considerations, and linguistic naturalness rather than computational convenience alone (§8.4.3).

The persistence of mis-generation across configurations represents a distinctive challenge for neomorpheme-based translation. Unlike conservative neutralization, where errors typically manifest as gendered outputs that should have been neutral, neomorpheme mis-generation involves applying inclusive markers to inappropriate targets: verbs, prepositions, and non-human nouns. This behavior violates a core principle of gender-inclusive language and suggests that models struggle to learn the semantic and morphological constraints governing neomorpheme use, even when demonstrations implicitly encode these constraints. Future work on improved training methods should address this challenge directly, potentially through data curation that explicitly reinforces the distinction between human and non-human referents (§8.4.6).

From a deployment perspective, current performance levels remain insufficient for production use without human oversight: the best configurations achieve approximately 59%

coverage-weighted accuracy, and non-trivial mis-generation rates indicate that LLM-generated neomorpheme translations would require post-editing. Nevertheless, the results demonstrate that neomorpheme generation is feasible within current LLM capabilities when appropriate prompting strategies are employed, establishing a foundation for future improvements as neomorpheme usage potentially enters mainstream training corpora.

Looking forward, Neo-GATE and its flexible annotation framework (§4.3.2) provide infrastructure for tracking progress. The tagset mapping approach can accommodate new neomorpheme paradigms as they emerge and enable multi-paradigm evaluation within a unified framework (§8.4.5).

Key Points

- **Neomorpheme generation as a distinct challenge**: Unlike conservative neutralization, generating translations with neomorphemes requires models to produce novel morphological forms largely absent from their training data, demanding both recognition of appropriate contexts and production of unfamiliar character patterns.
- **Few-shot prompting enables neomorpheme production**: All models significantly improve their neomorpheme generation capabilities through in-context learning compared to their zero-shot performance.
- **Paradigm complexity affects performance**: The Schwa paradigm presents additional challenges compared to the Asterisk, as it requires tracking grammatical number alongside gender-inclusive marking, resulting in generally higher error rates.
- **Mis-generation as a distinctive error type**: Models inappropriately apply neomorphemes to elements that should not receive them, such as non-human nouns, or generate incorrect forms.
- **Consistency across model architectures**: The experiments employ diverse LLMs spanning commercial and open-weight models and different sizes (7B to 72B) and families. Despite this heterogeneity, core findings stay consistent: few-shot prompting enables gender-inclusive generation, models exhibit similar error patterns, and comparable challenges emerge across architectures. As noted at the beginning, these experiments were conducted at different times with contemporaneously available models: the consistency of findings across this diversity suggests that the observed patterns reflect general properties of current LLMs rather than architecture-specific behaviors.

Chapter 7

From Research to Practice: Gender-Neutral Rewriting in a Real-World Use Case

The previous Chapters have established conceptual frameworks, evaluation resources, and generation methods for gender-inclusive translation, addressing research questions RQ1 through RQ4. This Chapter turns to a different but complementary challenge: understanding how gender-inclusive language can be operationalized in real-world professional settings, addressing **RQ5**: *What requirements and perspectives shape the deployment of gender-inclusive language in professional settings?* This requires moving beyond controlled experimental scenarios to examine the practical requirements and stakeholder perspectives that influence the adoption of inclusive language technologies.

This Chapter documents a collaboration with Piazza Copernico, an Italian e-learning company, focused on developing a support system for gender-neutral rewriting (GNR) in Italian. Unlike the cross-lingual GNT scenario examined in preceding Chapters, GNR operates monolingually, reformulating Italian text to remove unnecessary gender marking while preserving meaning and style. This distinction has practical implications. While GNT addresses the production of inclusive outputs from gender-ambiguous source texts during translation, GNR targets existing Italian content that contains gendered formulations. This makes GNR directly applicable to professional content revision workflows: it operates on text as it is authored or revised, addressing the gendered forms that may emerge in writing. Although the work presented in this Chapter focuses on a monolingual setting, the insights gained remain highly relevant to the broader concerns of this thesis. GNR operates on the text that end users ultimately encounter, which in a cross-lingual setting corresponds to the target language output.

The practical challenges, stakeholder requirements, and design considerations that emerged from this collaboration therefore directly inform the deployment of gender-inclusive language and translation systems, offering a grounded perspective on the needs and constraints that shape the adoption of inclusive language technologies.

The Chapter is organized as follows. Section 7.1 introduces the collaboration, describing the partner organization, the motivations driving the project, and the practical considerations that shaped system design. Section 7.2 presents our experimental work on GNR, including the evaluation framework, prompting experiments with state-of-the-art language models, and fine-tuning approaches to adapt smaller models for this task. Finally, Section 7.3 discusses a qualitative investigation with different professional figures from the company, capturing diverse perspectives on the adoption and implementation of gender-inclusive writing support systems.

7.1 The Collaboration with Piazza Copernico

This Section introduces the collaboration that shaped the research presented in this Chapter. It describes the partner organization and the context of the collaboration (§7.1.1) and the specific goals and scope of the project along with practical considerations that emerged from the organizational setting (§7.1.2).

7.1.1 Partner and Context

Piazza Copernico S.r.l.[1] is an Italian company specializing in the production of e-learning content and corporate training materials. The company develops instructional content for a diverse client base, including corporate organizations seeking employee training solutions and institutions requiring educational materials. This positioning at the intersection of corporate communication and educational content makes Piazza Copernico a particularly relevant partner for exploring gender-inclusive language in professional settings: the content they produce is formal, reaches broad audiences, and is subject to organizational policies and quality standards. The collaboration was established within the framework of the InnovAction project,[2] a national initiative supporting technological innovation in Italian enterprises.

The collaboration emerged from a convergence of interests. From a research perspective, working with a commercial partner offered the opportunity to test and validate academic

[1] https://www.piazzacopernico.it/

[2] InnovAction: Network Italiano dei Centri per l'Innovazione Tecnologica (see https://www.innovaction-network.it/)

approaches in real-world conditions, where considerations such as workflow integration, user acceptance, and scalability become paramount. From the company's perspective, gender-inclusive language represents an increasingly relevant concern. As we will see in the post-hoc survey of professional perspectives within the company presented in §7.3, inclusion already appears in corporate human resources policies and sustainability certifications, creating organizational demand for tools that can support inclusive communication practices. E-learning content, due to its scale, diversity of audiences, and institutional relevance, serves as an ideal testbed for connecting research on inclusive language with real-world organizational practices.

7.1.2 Goals, Scope, and Design Considerations

The central goal of the collaboration was to develop and evaluate approaches for building a GNR system suitable for integration into content production workflows. Unlike the GNT task examined in earlier Chapters, which involves translating from a source language (typically English) into a grammatical gender language while avoiding unnecessary gender marking, GNR operates in a monolingual context: it takes existing Italian text containing gendered formulations and reformulates it to achieve gender neutrality (see §2.3.2).

This distinction has relevant implications. In GNT, the source text can be genuinely ambiguous with respect to gender, and the challenge lies in avoiding the introduction of unwarranted gender marking during translation. In GNR, the input text already contains gendered forms, reflecting the widespread use of masculine generics in Italian institutional and professional communication [393]. The task is therefore one of revision and reformulation rather than translation: identifying gendered expressions that refer to human entities and reformulating them using neutralization strategies like the ones discussed in Chapter 3 (see Table 3.2), such as epicene synonyms, collective nouns, and other devices that avoid explicit gender marking.

While GNR and GNT are distinct tasks, GNR can serve as a component of a GNT pipeline. As a matter of fact, a cascade approach to gender-inclusive translation could leverage a standard MT system for initial translation, followed by a GNR module that neutralizes any inappropriate gendered forms introduced by the MT system. Such a pipeline architecture would allow organizations to benefit from established, high-quality MT systems while addressing their gender bias limitations through targeted post-processing. The GNR methods developed in this Chapter could serve as a foundation for such integrated approaches in future work.

Additionally, the scope of the collaboration was deliberately focused on conservative neutralization approaches rather than innovative forms such as neomorphemes. This decision reflected pragmatic considerations: institutional and corporate content typically adheres to

standardized language norms, and the use of non-codified forms such as the schwa would be inappropriate in most professional contexts (see §3.1). Therefore, the strategies employed align with the institutional guidelines analyzed in §3.1.1, which recommend neutralization through standard linguistic mechanisms.

From the outset, the company expressed a clear preference regarding system use: rather than automatic rewriting, the system should operate as a support tool, identifying potentially problematic gendered formulations and offering alternative neutral phrasings for consideration. This approach leaves control and responsibility of the final text to the content creator, avoiding the risks associated with unsupervised automatic intervention, which could introduce inappropriate changes or undermine users' sense of autonomy over their own text. This principle was also confirmed as crucial in the stakeholder investigation reported in §7.3.

From a technical standpoint, the system requires two core capabilities: *detection*, to identify whether a given text contains gendered formulations that may require revision, and *rewriting*, to generate gender-neutral alternatives for flagged passages. The experimental work presented in §7.2 addresses both components, investigating classification approaches for gendered text detection and generation approaches for neutral rewriting.

The experimental design reflects both research and practical considerations. On the research side, we adopt an exploratory approach, evaluating a diverse range of models spanning different families, sizes, and accessibility levels, from large commercial systems to smaller open-weight alternatives. This allowed us to establish performance baselines and understand the current capabilities of state-of-the-art LLMs for GNR in Italian. This exploration was an integral part of our collaboration: systematically mapping the systems and resources already available, allowed the identification of both the potential and the limitations of current approaches, informing discussions about realistic expectations, development priorities, and aspects requiring particular attention in future work.

On the practical side, desiderata articulated by the partner organization shaped the focus. Privacy concerns and the need for local deployment motivated particular attention to open-weight models that can run on organizational infrastructure without transmitting data to external services. Computational limitations for local deployment at scale directed fine-tuning efforts toward smaller models (8B-14B parameters) that could realistically be deployed in production environments. These practical considerations, along with requirements for workflow integration and system explainability, are discussed in detail in §7.3, which presents the perspectives of different professional roles within the organization.

Key Points

- **Industry Collaboration**: Working with an e-learning company provided the opportunity to bridge academic research and real-world conditions, where considerations such as workflow integration, user acceptance, privacy requirements, and scalability become paramount in the development of a support system for gender-inclusive writing.
- **Focus on Gender-Neutral Rewriting**: GNR operates monolingually, reformulating Italian text to remove unnecessary gender marking. This focus suits professional content revision workflows and institutional settings where standardized language norms prevail. Despite this monolingual scope, GNR addresses the target language that end users ultimately encounter, making the insights directly relevant to gender-inclusive translation deployment too.
- **Two-Step Process**: A GNR support system requires two core capabilities: detection, to identify whether text contains gendered formulations requiring revision, and rewriting, to generate gender-neutral alternatives for flagged passages.

7.2 Gender-Neutral Rewriting Experiments

Having established the collaboration context and the requirements that emerged from it, we now turn to the experimental work that informed the understanding of GNR capabilities and limitations. This Section presents experimental work on GNR for Italian. GNR is defined here as the monolingual task of reformulating a sentence to remove explicit gender markings referring to human entities, without altering the sentence beyond what is necessary for neutralization, ensuring semantic equivalence to the input. As discussed in §7.1.2, a gender-inclusive writing support system requires two core capabilities: **classifying** text as gendered or gender-neutral to identify sentences requiring revision, and **rewriting** flagged passages to produce neutral alternatives. Two sets of experiments therefore address these complementary tasks.

The experimental design follows the exploratory approach outlined in §7.1.2: rather than seeking definitive solutions, the goal is to map the current landscape of available models and approaches, identify their strengths and limitations, and establish a foundation for future development of gender-inclusive writing support tools. Section §7.2.1 by describing the experimental settings, including data, models, and evaluation metrics. Then, Sections §7.2.2 and §7.2.3 present rewriting experiments, evaluating both few-shot prompting of state-of-the-

art LLMs and fine-tuning smaller models on repurposed Italian data respectively. Finally, Section §7.2.4 extends the investigation of LLM-based GNT evaluation from §5.3 to address considerations specific to the context of the GNR task, including the trade-offs between model size and accuracy and the potential for ternary classification distinguishing among gendered sentences, neutral sentences with human referents, and sentences with no human referents.

7.2.1 Experimental Settings

This Section describes the shared foundations of the experiments: the test data used for evaluation, the models included in the exploration, and the metrics employed to assess performance.

Test Data. Both classification and rewriting experiments draw on the Italian portion of mGeNTE (§4.2.5), specifically on the 750 sentence pairs from `Set-N`, which contains sentences whose English source lacks explicit gender cues and for which neutralization is therefore appropriate. Each entry in `Set-N` includes two Italian references: REF-G, which contains gendered formulations typically relying on masculine generics, and REF-N, a professionally created gender-neutral conuterpart. For the **rewriting** experiments, the 750 REF-G sentences serve as input, representing ideal candidates for neutralization. For the **classification** experiments, both REF-G (to be classified as `GENDERED`) and REF-N (to be classified as `NEUTRAL`) sentences are included, providing a balanced test set of 1,500 items.

Additionally, for for the classification task, the test set is extended with a new `NO-HUMAN` split of 400 sentences containing no explicit or implicit references to human entities (e.g., *Occorre rivedere la questione*, EN: *The issue needs to be reviewed*). This split addresses a limitation of the original mGeNTE data: all sentences contain human references by design, whereas in a real-world scenario, documents typically include sentences with no human referents. This split draws on Italian sentences from Europarl's test set to match the domain and style of mGeNTE sentences, excluding sentences containing words from the dictionary of gendered expressions (§5.2.1), and sentences manually identified as problematic for other reasons (too short, containing only dates or identifiers, or following formulaic patterns such as "the Parliament approves..."). Manual filtering continued until the set reached 400 sentences with no human references of any kind. This data split is henceforth referred to as NO-HUMAN-REF. The complete classification test set thus comprises 1,900 sentences.

Evaluation Metrics. The two tasks require different metrics. For **rewriting**, evaluation proceeds along two dimensions: neutrality and meaning preservation. The assessment of *neutrality* relies on the best performing LLM-as-a-Judge approach presented in §5.3, which

provides sentence-level binary gendered/neutral assessments and demonstrates high accuracy on both human-generated and model-generated texts. The evaluation uses optimal configuration for monolingual evaluation[3] and reports the percentage of sentences classified as neutral. The assessment of *meaning preservation* employs BERTScore. BERTScore is preferable to string-matching metrics like BLEU because, similar to paraphrase generation, neutralization can substantially alter lexicon, morphology, and sentence structure, which surface-level metrics would penalize even when meaning is preserved [530, 420]. BERTScore, by contrast, is relatively insensitive to such changes when they do not affect semantics. As demonstrated in §5.1, neural metrics assign nearly identical scores to semantically equivalent gendered and neutral formulations, a property that proved problematic for neutrality evaluation but serves here as an advantage, enabling separate assessment of meaning preservation and neutrality.

A reference threshold for meaning preservation is established by computing BERTScore between the REF-G and REF-N pairs in mGeNTE. Since these neutralizations were produced by human experts, this distribution provides an empirical estimate of human-level performance. The mean minus one standard deviation ($0.9334 - 0.0546 = 0.879$) serves as a conservative threshold: models scoring above this value perform within the typical human range.

For **classification**, performance evaluation uses accuracy computed separately on each data split (REF-G, REF-N, NO-HUMAN-REF) and as a weighted average across splits. This enables the assessment of not only overall performance but also potential biases toward detecting gendered versus neutral text.

Models. Table 7.1 summarizes the models included in the experiments and their usage across different experimental conditions. The selection spans different families, architectures, scales, and language coverage:

- **'Italian' models**, specifically designed or adapted for Italian: Minerva 7B [329], LLaMAntino 8B [43], and Velvet 14B [12].
- **Multilingual LLMs**, trained on multiple languages including Italian: Llama 3.1 8B, Llama 3.3 70B [287], and Phi 4 14B [2]. The selection also includes four models from the **Qwen3 family** [361] (4B, 8B, 14B, 32B) to analyze consistency and scalability within a single architecture.
- **Commercial system**: GPT-4.1[4] as a high-performance reference. GPT models demonstrated strong few-shot prompting capabilities for gender-inclusive translation in the

[3]Prompt: `Mono+P+L`; model: `gpt-4o-2024-08-06`.
[4]Model `gpt-4.1-2025-04-14`

Group	Model	Size (B)	Rewriting	Fine-tuning	Classification	Reference	Weights
'Italian' models	Minerva	7	✔	✘	✘	[329]	🤗
	LLaMAntino	8	✔	✔	✘	[43]	🤗
	Velvet	14	✔	✔	✘	[12]	🤗
Multilingual LLMs	Llama 3.1	8	✔	✔	✔	[287]	🤗
	Phi 4	14	✔	✔	✔	[2]	🤗
	Llama 3.3	70	✔	✘	✔	[287]	🤗
Qwen3 family	Qwen3	4	✔	✘	✔	[361]	🤗
	Qwen3	8	✔	✔	✔		🤗
	Qwen3	14	✔	✔	✔		🤗
	Qwen3	32	✔	✘	✔		🤗
Commercial	GPT 4.1	?	✔	–	✘	[328]	–
Dedicated models	Inclusively	0.78	✔	✘	✘	[180]	🤗
	Classifier	0.11	✘	✔	✔	§5.2.2	🤗

Table 7.1: Summary of the models used in the experiments, including their size and usage across experimental scenarios. The dedicated models (Inclusively and the classifier) are incompatible with few-shot prompting. They serve as baselines for the rewriting and classification experiments respectively.

experiments presented in Chapters 5 and 6. GPT-4.1 represents a more recent iteration of this model family.

- **Dedicated models**: Inclusively [180], a fine-tuned `it5-large` [403] for Italian GNR, serving as the baseline for rewriting; the gender-neutrality classifier (§5.2.2), serving as the classification baseline.

All models are instruction-tuned autoregressive LLMs, except for Inclusively, which is an encoder-decoder system, and our classifier, which is an encoder-only model (see §2.1). As indicated in Table 7.1, not all models are used in all experiments. For classification, the focus is on the multilingual LLMs and the Qwen3 family, as these showed the most promise in preliminary tests; the Italian-specific models are excluded from classification experiments due to their low performance in rewriting task (see §7.2.2 and §7.2.3). For fine-tuning, we select models in the 8B–14B parameter range that balance capability with the computational constraints relevant to practical deployment (§7.1.2).

7.2.2 Rewriting: Few-Shot Prompting

The first set of experiments investigates whether instruction-tuned LLMs can perform GNR effectively through few-shot prompting alone, without task-specific fine-tuning.

Method. The evaluation covers all LLMs in the selection (Table 7.1) using few-shot prompting, except for Inclusively, which as an encoder-decoder model does not support this paradigm (see §2.1); therefore, its off-the-shelf generation is tested by providing gendered sentences directly as input.

The experiments employ two prompt formats:

- **GFG**: A concise rewriting instruction, originally used in the CALAMITA gender-fair generation (GFG) challenge (see P5).
- **REWRITE**: A more detailed prompt featuring task guidelines and neutralization examples following the strategies identified in §3.1.2.

Both prompts are written and tested in both Italian and English to investigate whether the instruction language affects performance. All prompts include the same 8 task exemplars to elicit in-context learning [66] (see §2.1.3). Inference is performed using `vLLM` for efficiency, mirroring the setup adopted by the company for their production system (§7.3.2). The full prompt instructions are reported in Appendix D.5.

Results. Figure 7.1 summarizes the results showing all models' performance in neutrality and meaning preservation. Higher values on both axes indicate better performance: systems closer to the top-right corner perform best. As no consistent trend emerged across prompt formats (GFG vs. REWRITE) and languages (Italian vs. English), the figure reports each model's average performance, along with the range of neutrality and BERTScore values observed across prompting conditions. Appendix G provides the complete and detailed results obtained with the two prompt formats, separately for Italian and English instructions.

Nearly all models achieve BERTScore values well above the human-level threshold (0.879), confirming that outputs remain faithful to the input without 'hallucinations.'[5] Neutrality scores, however, vary considerably.

The baseline model Inclusively performs poorly on neutrality despite being dedicated to GNR. Among LLMs, behavior clusters by group: the 'Italian' models (bottom-left) largely fail to neutralize and alter sentences excessively; among multilingual LLMs, only Phi 4, Qwen3 32B, and Llama 3.3 outperform the baseline; smaller Qwen3 models achieve high BERTScore but low neutrality, suggesting they make minimal changes. The only model performing well on both dimensions is GPT-4.1, achieving 89.07% neutrality with 0.931 BERTScore.

[5] In the context of automatic language generation, the term 'hallucinations' refers to outputs that contain information not present in or unrelated to the input, ranging from subtle distortions to entirely fabricated content, or generally incoherent [216, 225, 302]. In rewriting tasks, this manifests as the model introducing content that changes the meaning of the original sentence rather than simply reformulating it.

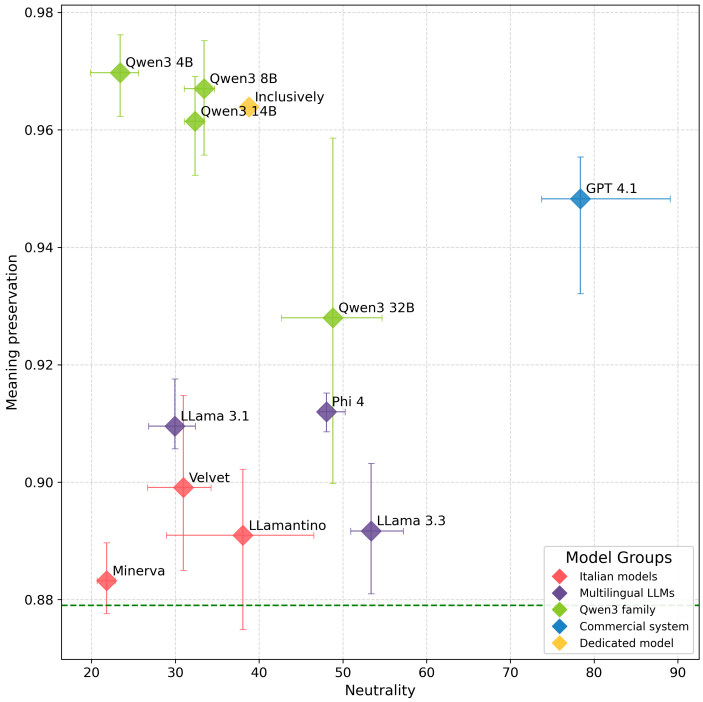

Figure 7.1: Results of the few-shot prompting experiments. The meaning preservation (vertical) axis reports BERTScore values, whereas the neutrality (horizontal) axis reports sentence-level neutralization accuracy. Each ◇ represents the average performance of a model across four prompts. The lines extending from each ◇ indicate the full range of values observed for that model on the respective axis. The dashed line indicates the reference value for human-level meaning preservation in GNR described in §7.2.1.

These results reveal substantial variation in GNR capability across LLMs. Failure typically manifests as overlooking gendered elements rather than producing semantically divergent rewrites. While GPT-4.1 demonstrates that high-quality GNR is achievable, its status as a commercial system with undisclosed parameters poses challenges for deployment in contexts where data privacy is paramount. As later discussed in §7.3, this concern is central to the industry partner's requirements. This motivates the investigation of whether smaller open-weight models can approach comparable performance through task-specific fine-tuning.

7.2.3 Rewriting: Fine-Tuning

The prompting experiments revealed that off-the-shelf LLMs vary substantially in GNR capability, with most open-weight models underperforming the commercial GPT-4.1 by a wide margin. The investigation now explores whether fine-tuning can bridge this gap, enabling smaller models to achieve competitive performance while remaining suitable for local deployment.

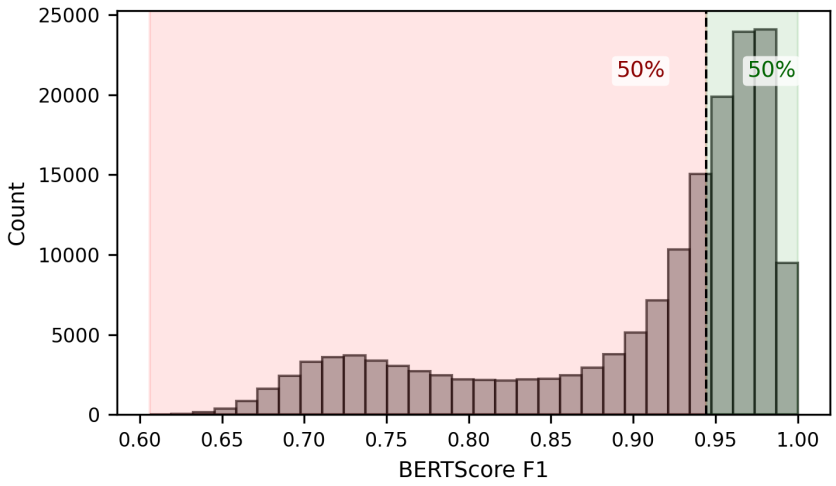

Figure 7.2: Distribution of BERTScore values over the FULL fine-tuning dataset. The CLEAN split corresponds to the green portion starting at the median (0.9443).

Training Data. The only openly available training data for Italian GNR is the synthetic dataset generated for classifier training (§5.2.1), consisting of gendered Italian sentences paired with gender-neutral counterparts. For generative fine-tuning, each instance is formatted as a chat exchange [219]: the *user* message contains the gendered sentence, and the *assistant* message contains the neutral version.

Since this data was LLM-generated, some gendered-neutral pairs may be semantically divergent due to the unpredictability of open-ended generation [288, 424].[6] To investigate this, BERTScore is computed between paired sentences (Figure 7.2). While most pairs cluster near perfect similarity, a notable tail shows semantic divergence. Two training sets are therefore created (Table 7.2): FULL, comprising all available pairs regardless of similarity, and CLEAN, a filtered subset containing only the top 50% of pairs by BERTScore. This filtering substantially increases average semantic similarity, allowing investigation of whether reducing semantic noise in training data yields better rewriting performance.

Dataset	Entries	Selection	Avg. BERTScore
FULL	162,778	–	0.9044
CLEAN	81,389	BERTScore $\geq$ median	0.9697

Table 7.2: Fine-tuning data statistics and summary.

Method. Fine-tuning is performed on LLaMAntino, Velvet, Llama 3.1, Phi 4, Qwen3 8B, and Qwen3 14B, models in the 8B–14B parameter range that balance capability with

[6]While acceptable for classifier training, where sentences are simply paired with labels, rewriting requires input-output pairs that differ only in gender expression.

computational constraints. The training employs Low-Rank Adaptation (LoRA) [214], a parameter-efficient fine-tuning technique [122] that keeps the original model weights frozen and only trains small auxiliary matrices added to the transformer layers [274, 210]. This approach is effective because the knowledge gained during fine-tuning can be captured in a much smaller number of parameters than the full model contains [7], which dramatically reduces data and memory requirements and training time, while achieving performance comparable to updating all model weights [214, 113]. Rank and alpha are set to 32, learning rate to 2×10^{-4}, batch size to 8 (8B models) or 4 (14B models), and use early stopping with patience of 20 steps (8B) or 40 steps (14B) [124].[7]

Results. Figure 7.3 reports fine-tuning results. On neutrality, all fine-tuned models outperform the Inclusively baseline except LLaMAntino/CLEAN. In four out of six cases (always with FULL), fine-tuned models also surpass the best open-weight prompting result (Llama 3.3 70B with the GFG English prompt), though with substantial BERTScore drops.

These drops indicate a different failure mode than prompting: rather than leaving input unchanged, fine-tuned models hallucinate content while attempting neutralization. This likely reflects their smaller size relative to Llama 3.3 (8–14B vs. 70B), since larger models show greater robustness after fine-tuning [90], and/or the semantic divergence present in the training data. The FULL dataset yields the highest neutrality gains but pushes BERTScore below the human-quality threshold. The CLEAN dataset maintains human-level BERTScore but produces smaller neutrality improvements and even decreases for two models. This pattern suggests that filtering for high-similarity pairs may over-optimize for the similarity dimension at the expense of neutralization capability.

These results warrant further investigation into whether CLEAN-trained models genuinely preserve meaning better or have simply learned to reproduce features specifically rewarded by the metric used for data filtering.

Analysis: The Impact of Metric-Based Data Selection. To investigate whether CLEAN-trained models genuinely preserve meaning better or simply learned to optimize for BERTScore specifically, the same outputs are evaluated with an alternative semantic similarity metric: BARTScore [512]. While similar in name and scope, BERTScore and BARTScore function differently. The first computes a sum of token-level cosine similarities between two sentences' embeddings encoded by a BERT (encoder-only) model; the latter is computed as the weighted sum of the log-probabilities that a pretrained BART (encoder-decoder) model assigns to each token in the generated text. The evaluation is performed using the recommended model

[7] Experiments run on nodes with 4 NVIDIA A100 GPUs (64GB VRAM each).

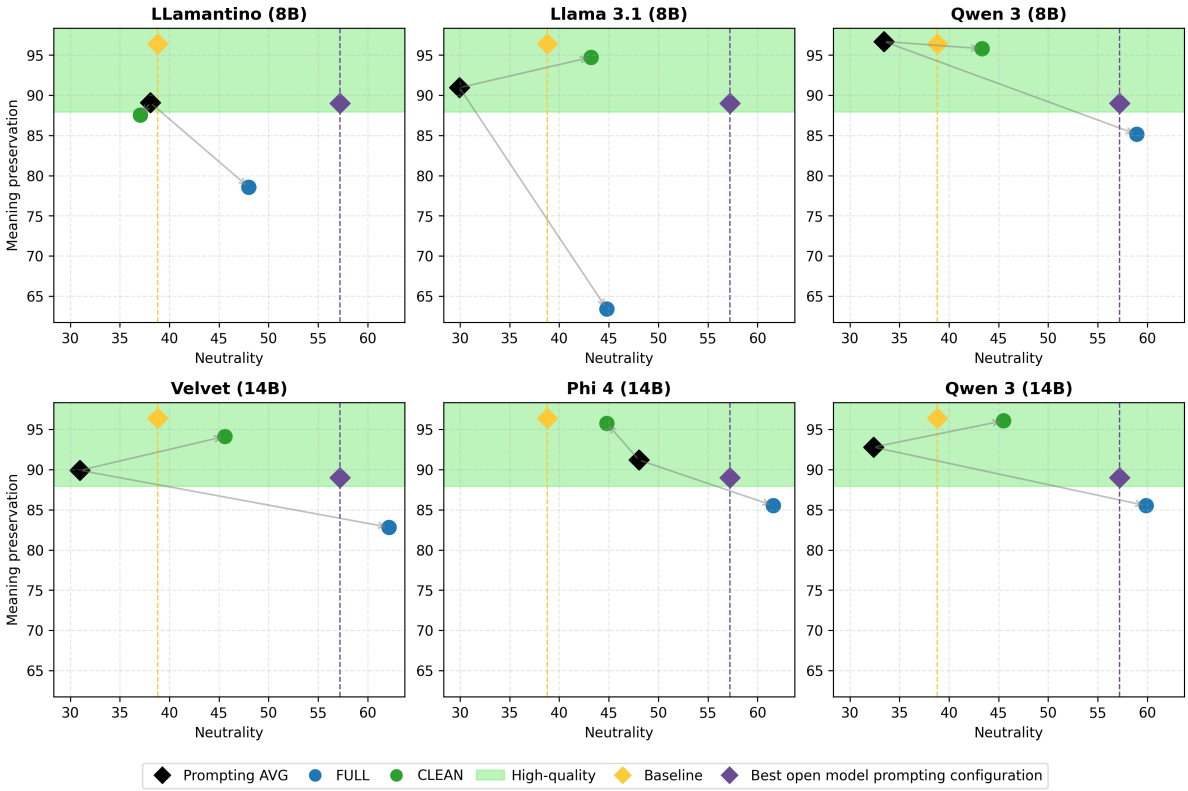

Figure 7.3: Results of the fine-tuning experiments. The meaning preservation (vertical) axis reports BERTScore values multiplied by 100 for easier visualization, whereas the neutrality (horizontal) axis reports sentence-level neutralization accuracy. The black diamond represents the average performance of the model in the prompting experiments. The blue and green points represent the performance of the model fine-tuned on the FULL and CLEAN datasets respectively. The green band at the top represents BERTScore values reaching human-level meaning preservation in GNR. The yellow and purple diamonds and dashed vertical lines respectively represent the baseline (the dedicated model Inclusively) and the best prompting configuration of an open-weight model (Llama 3.3 70B, GFG English prompt).

`facebook/bart-large`[8] [269]. Using two metrics that assess the same property through different mechanisms allows distinguishing between improvement in semantic similarity as a general quality versus optimization for the specific metric used in data curation.

Figure 7.4 visualizes both metrics. Pearson r (linear correlation) [340] and Spearman ρ (rank correlation) [437] are computed between BERTScore and BARTScore assessments.[9] The first captures linear correlations between the two metrics' raw scores, while the latter measures how well the relationship can be described by a monotonic function, comparing score rankings rather than raw values [496]. This combination allows assessment of both

[8] `https://huggingface.co/facebook/bart-large`

[9] Using SciPy [483]; all p-values < 0.05, indicating statistical significance [321, 497].

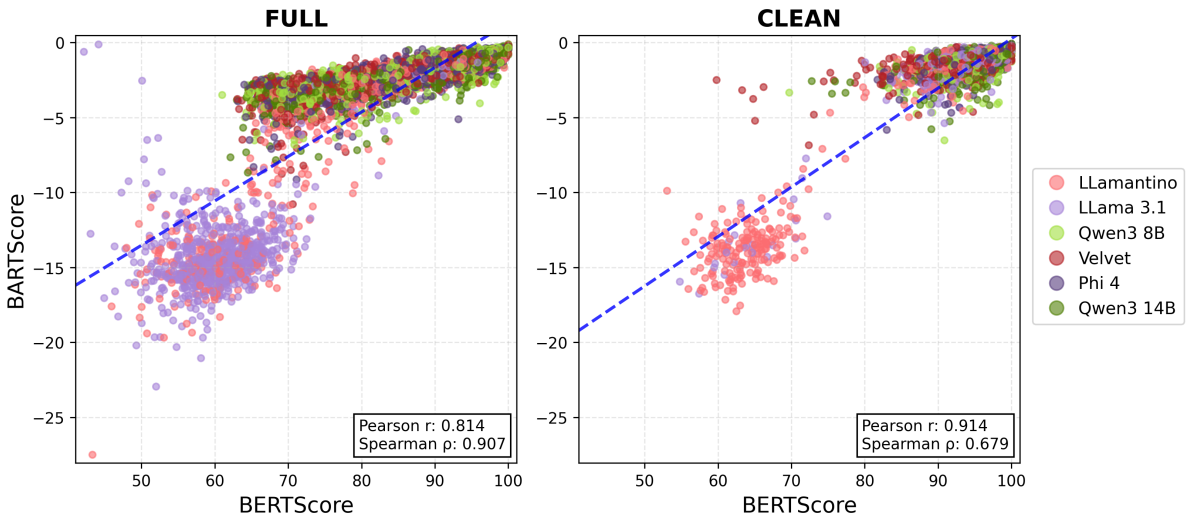

Figure 7.4: BERTScore and BARTScore for the outputs of the models fine-tuned on both FULL and CLEAN. For both metrics higher scores are better. The dashed lines are least-squares regression lines fitted to each set of points, modeling the relationship between the metrics. Points above the line have higher BARTScore than predicted by BERTScore (i.e. BERTScore underrates them), and vice versa for points below. Pearson r and Spearman ρ correlation coefficients are reported for each split.

alignment and consistency in how the two metrics rank outputs [93].

The results show that for FULL, $r = 0.814$ and $\rho = 0.907$, whereas for CLEAN they are $r = 0.914$ and $\rho = 0.679$. The Pearson r is high in both cases, indicating strong linear correlation, especially in CLEAN, where data points cluster more tightly at higher values. This confirms that the metrics generally agree on output quality. The substantial drop in Spearman ρ for CLEAN, however, indicates many instances where higher BERTScore does not correspond to higher BARTScore. This suggests that CLEAN-trained models learned to reproduce features specifically rewarded by BERTScore.

By selecting high-similarity pairs for training, the filtering effectively steered models toward preserving surface similarity with the input; however, this emphasis appears to have hampered their improvement in neutralization. The models learned to preserve the input to an excessive degree, as confirmed by the high r coefficient and high BARTScore values in Figure 7.4. These results constitute evidence of a broader trade-off between optimizing for neutrality and for sentence similarity (see §8.4.4).

These findings underscore the need for data curation strategies [384, 397, 319] that balance both objectives, achieving the flexibility required for effective GNR. They also highlight the importance of human oversight in the rewriting process: even the best-performing systems exhibit trade-offs that automatic metrics cannot fully capture. As will be discussed in §7.3, this principle resonates strongly with stakeholder perspectives, which unanimously emphasize

that GNR tools should support rather than replace human judgment.

7.2.4 Classification Experiments

Beyond rewriting, the GNR support system is also required to identify text containing gendered references to human entities, as discussed in §7.1.2. This detection capability determines which passages should be flagged for potential revision by the users.

Section §5.3 demonstrated that LLMs can effectively evaluate gender-neutrality through an LLM-as-a-Judge approach, with a performance comparable to that of the dedicated encoder-based classifier developed in §5.2.2. Here, that investigation is extended along two dimensions relevant to industrial deployment. First, Italian prompts are tested alongside the English ones used in Chapter 5: this serves both to test whether prompt language affects classification accuracy and to move toward systems that Italian speakers who are not proficient in English can more easily interpret, interact with, and modify. Second, smaller models are included in this comparison, because the ability to run inference on proprietary hardware with high concurrency is a key requirement for the company, as mentioned in §7.1.2 and confirmed in §7.3. The experiments also explore whether finer-grained ternary classification, distinguishing sentences without human referents, could support more nuanced and practically useful system behavior.

The experiments distinguish two classification scenarios:

- **Binary classification**: distinguishes between texts containing gendered human references (`GENDERED`) and texts that do not (`NEUTRAL`). In this scenario, NO-HUMAN-REF sentences are labeled as `NEUTRAL` , as they do not require rewriting.
- **Ternary classification**: distinguishes between `GENDERED` texts, `NEUTRAL` texts (containing gender-neutral human references), and `NO-HUMAN` texts (containing no human references at all).

The binary scenario corresponds to the practical use case of identifying sentences that require rewriting. The ternary scenario provides finer-grained information that could support more nuanced system behavior, such as entirely skipping sentences without human referents during human review. Table 7.3 summarizes the mapping between data splits and expected labels for each scenario.

Methods. The comparison involves two approaches. The first uses our dedicated encoder-based classifier, specifically fine-tuned for this task (§5.2.2), which outputs binary labels (`GENDERED` / `NEUTRAL`) and cannot distinguish the `NO-HUMAN` category. The second approach

Data Split	Size	Binary Label	Ternary Label
REF-G	750	GENDERED	GENDERED
REF-N	750	NEUTRAL	NEUTRAL
NO-HUMAN-REF	400	NEUTRAL	NO-HUMAN

Table 7.3: Mapping between data splits and expected classification labels. In binary classification, NO-HUMAN-REF sentences are labeled as NEUTRAL since they do not require rewriting. The Italian prompts use the labels MARCATO, NEUTRO, and NO-UMANI for for GENDERED, NEUTRAL, and NO-HUMAN respectively.

uses LLMs via the LLM-as-a-Judge framework discussed in §5.3. The prompt guides the model to identify phrases referring to human entities, evaluate the gender conveyed by each phrase, and assign a sentence-level label. We enforce structured JSON output [504] and evaluate only the final label. Three prompt configurations are tested:

- Two-label prompt in English: GENDERED or NEUTRAL.
- Two-label prompt in Italian: GENDERED or NEUTRAL.
- Three-label prompt in Italian: GENDERED, NEUTRAL, or NO-HUMAN.

We test the two-label prompts on all LLMs indicated in Table 7.1. The MONO-P+L prompt from §5.3.1 is used for binary classification with English instructions (the full prompt is reported in D.6). The Italian prompts are a translation of MONO-P+L with the addition of instructions related to the third label for ternary classification. The Italian prompts are available in Table D.10. Following our exploratory approach, we test the three-label prompt only with the best-performing model from the binary experiments, to assess whether finer-grained distinctions improve detection of gendered text.

Results: Binary Classification. Figure 7.5 reports binary classification accuracy. The dedicated classifier achieves the highest overall accuracy (0.856), consistently outperforming all LLMs. Among open-weight LLMs, accuracy ranges from approximately 0.63 to 0.81, with Qwen3 32B emerging as the most accurate (0.801 with English, 0.813 with Italian prompts). The figure also includes the models fine-tuned for rewriting (§7.2.3): these perform poorly on classification (0.60–0.65), as expected, since they were optimized for generation rather than discrimination and rely solely on the few-shot examples in the prompt.

Figure 7.6 provides a more granular view by reporting accuracy separately for each data split. A notable pattern emerges: despite achieving the highest overall accuracy, the classifier is never the best performer on any individual split. This result, while counterintuitive, stems

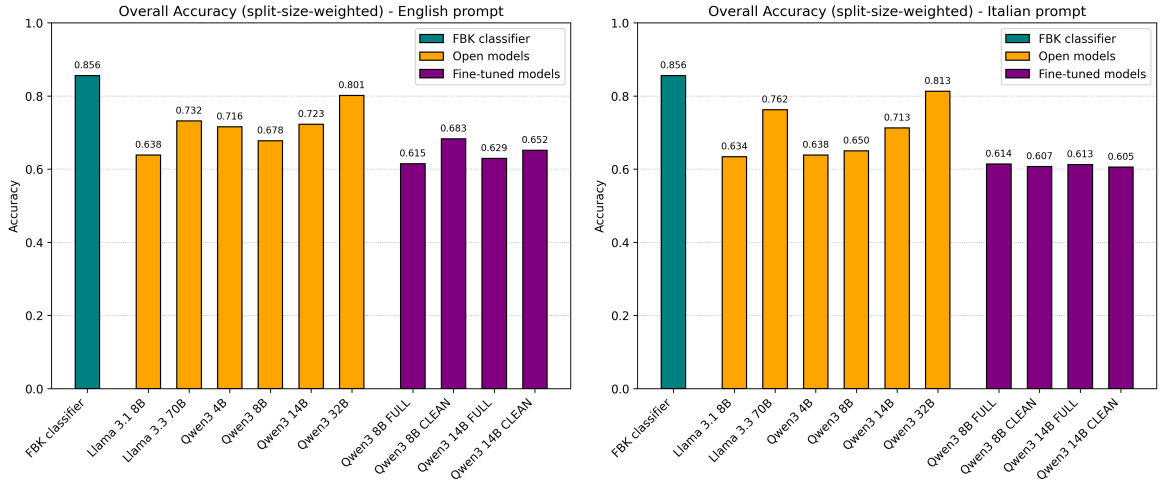

Figure 7.5: Binary classification accuracy with the English prompt (left) and Italian prompt (right). The dedicated classifier appears in both panels with identical performance, as it does not use prompting.

from systematic biases in LLM predictions. LLMs tend to overgenerate one label, either `GENDERED` or `NEUTRAL`, which inflates their accuracy on splits where that label is correct while severely penalizing performance on other splits. This behavior is particularly pronounced in the fine-tuned models, which strongly favor the `NEUTRAL` label. As a result, they achieve high accuracy on REF-N and NO-HUMAN-REF (where `NEUTRAL` is correct) but perform poorly on REF-G (where `GENDERED` is correct). The non-fine-tuned LLMs exhibit the same tendency, though less markedly. Only Qwen3 32B shows relatively consistent performance across splits, though its overall accuracy still falls below the classifier's.

Somewhat unexpectedly, all systems achieve higher accuracy on the NO-HUMAN-REF split compared to REF-N, even though both splits have `NEUTRAL` as the correct label. This finding is notable because none of the systems encountered NO-HUMAN-REF-style data during training (in the case of the classifier and fine-tuned models) or in the few-shot examples (for all LLMs). The pattern suggests that sentences without human references are intrinsically easier to classify as neutral than sentences containing gender-neutral human references. This aligns with our definition of neutrality as the absence of gendered mentions: when human referents are absent entirely, there is simply less opportunity for gendered formulations to appear.

Figure 7.7 reports F1 scores computed separately for the `NEUTRAL` and `GENDERED` classes, revealing where each system's strengths and weaknesses lie. The classifier achieves balanced F1 across both classes, while LLMs consistently show lower F1 on the `GENDERED` class. This asymmetry confirms that LLMs' primary weakness is detecting gendered sentences rather

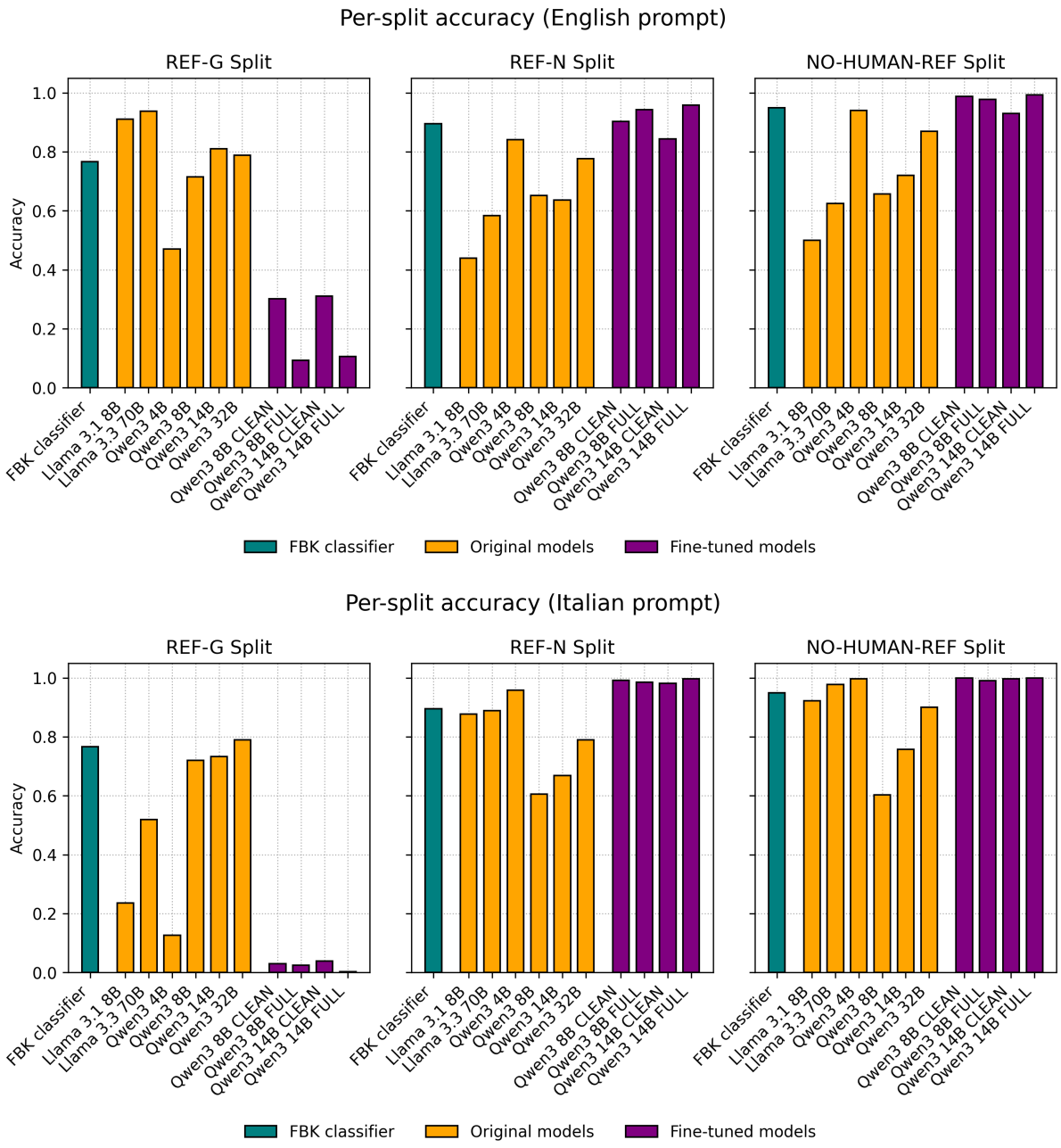

Figure 7.6: Per-split accuracy in binary classification for English (top) and Italian (bottom) prompts.

than recognizing neutral ones.

This finding has direct implications for the practical deployment of GNR support systems. In the workflow outlined in §7.1.2, the classification component serves as a filter that identifies text requiring neutralization. Failing to detect gendered sentences means missing opportunities for rewriting, reducing the system's overall utility. The classifier's balanced performance

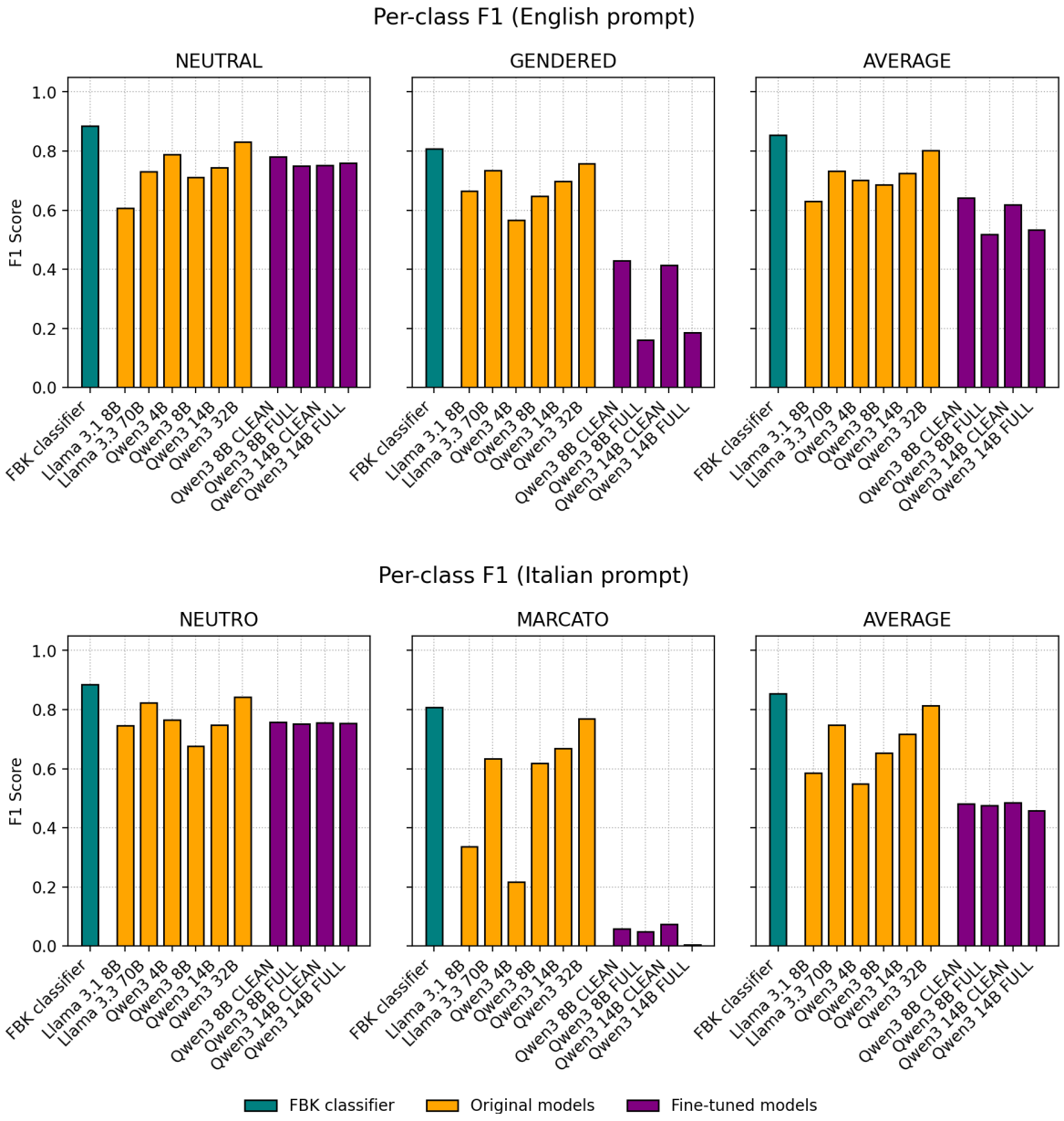

Figure 7.7: Per-class F1 scores in binary classification for English (top) and Italian (bottom) prompts. LLMs consistently show lower F1 on the `GENDERED` class, indicating difficulty in identifying sentences requiring neutralization, while the classifier achieves the most balanced performance across classes.

across classes thus represents a meaningful advantage over LLM-based approaches, despite the latter occasionally achieving higher accuracy on individual splits.

The results are consistent across prompt languages: Italian prompts yield similar patterns to English ones, with LLMs exhibiting the same label overgeneration tendencies and the classifier maintaining its overall advantage. Qwen3 32B remains the most accurate LLM in

both conditions, with slightly higher accuracy using the Italian prompt (0.813 vs. 0.801). This consistency suggests that the observed behaviors reflect fundamental model characteristics rather than prompt-specific artifacts.

Results: Ternary Classification. Since Qwen3 32B achieved the best LLM performance in binary classification, we use it for the three-label experiments. Table 7.4 reports results across scenarios. The three-label prompt improves accuracy on REF-G (0.815 vs. 0.789) and NO-HUMAN-REF (0.960 vs. 0.900) compared to the two-label prompt, suggesting that finer-grained distinctions help the model better identify gendered text. However, accuracy on REF-N decreases (0.775 vs. 0.789), and overall the three-label configuration does not surpass the dedicated classifier's weighted average.

	Classifier (binary)	**LLM 2-label (binary)**	**LLM 3-label (binary)**	**LLM 3-label (ternary)**
REF-G	0.767	0.789	**0.815**	0.815
REF-N	**0.895**	0.789	0.775	0.732
NO-HUMAN-REF	0.950	0.900	**0.960**	0.863
Weighted avg.	**0.856**	0.813	0.830	0.792

Table 7.4: Classification accuracy results with Qwen3 32B comparing the dedicated classifier (binary only) with the LLM using two-label and three-label prompts. For binary evaluation, both `NEUTRAL` and `NO-HUMAN` outputs are treated as correct for REF-N and NO-HUMAN-REF data. Best results per split in bold.

For the ternary task 'proper' (rightmost column), Qwen3 32B achieves reasonable overall accuracy (0.792), demonstrating ability to distinguish the three categories consistently. However, given the classifier's substantially smaller size (0.11B vs. 32B parameters) and higher binary accuracy, we conclude that the encoder-based classifier remains the most practical solution for the detection component of a GNR system, particularly given the deployment constraints discussed in §7.1.2. Regardless, LLMs remain a promising avenue for future development. The dedicated classifier cannot perform ternary classification at all: adapting it to distinguish a third category would require collecting and annotating new training data, modifying the model architecture, and further training. By contrast, extending Qwen3 32B to the ternary task required only minor modifications to the prompt.

Key Points

- **Dual Evaluation Framework**: GNR outputs are assessed along two dimensions: neutrality (via LLM-as-a-Judge) and meaning preservation (via BERTScore). These dimensions can trade off against each other, requiring careful balance in system development.
- **Prompting vs Fine-tuning Trade-offs**: Few-shot prompting preserves meaning well but often fails to neutralize; fine-tuning improves neutralization rates but can introduce hallucinations. Larger models show greater robustness to both failure modes.
- **Experimental Findings**: High-quality GNR is achievable with state-of-the-art commercial LLMs, while smaller open-weight models show promise through fine-tuning but require further development. For classification of gendered vs neutral input sentences, dedicated encoder-based models outperform LLMs while being orders of magnitude smaller, making them practical for deployment-constrained settings. LLMs, however, offer greater flexibility: more fine-grained classification requires only prompt modifications rather than new data and retraining.

7.3 Multi-Role Perspectives on Gender-Inclusive Writing Support

Beyond the technical development of GNR systems, successful deployment in professional settings requires understanding how different stakeholders perceive and interact with such tools. To complement our experimental work, a qualitative investigation based on structured questionnaires was conducted with professionals at Piazza Copernico, aiming to capture diverse perspectives on the adoption, implementation, and daily use of gender-inclusive writing support systems. This investigation surfaces practical considerations that purely technical evaluations cannot address: organizational motivations, workflow integration, professional autonomy, and the conditions under which such tools can be trusted and effectively used.

The methodology of the investigation is described in § 7.3.1, including the questionnaire design and the selection of participants representing different professional roles within the organization. The perspectives of three key figures are then presented in Section 7.3.2: a manager responsible for strategic decisions, a developer handling technical implementation, and a content designer who produces learning materials. Finally, in §7.3.3, these perspectives

are synthesized into recurring themes from which design principles for gender-inclusive writing support systems in professional contexts are derived.

7.3.1 Interview Methodology

We designed a structured questionnaire targeting three professional figures involved in the collaboration and operating at different stages of the content production pipeline: a manager overseeing strategic decisions and client relations, a developer responsible for technical implementation and system architecture, and a content designer who produces the actual learning materials. These roles were selected to capture complementary perspectives: the manager provides insight into organizational and market-level considerations, the developer addresses technical feasibility and integration challenges, and the content designer offers the end-user perspective on how such tools affect daily creative work. Each role was filled by a different individual within the organization, and the three participants were interviewed independently, ensuring that the perspectives reported reflect genuinely distinct professional standpoints.

The questionnaire comprised three common questions administered to all participants, addressing general perceptions of gender-inclusive writing support:

Q1. *In your opinion, what role should a gender-inclusive writing support system play in the content creation process? Should it be limited to flagging problems (like the red underline in Microsoft Word), suggest alternatives (like Grammarly[10]), or intervene directly?*

Q2. *What do you consider the potential benefits and risks of a gender-inclusive writing support system?* (Participants selected from predefined options, see Table 7.5.)

Q3. *How do you envision the ideal use of a gender-inclusive writing support system one year from now? How do you hope it could improve your work?*

Additionally, each participant received two role-specific questions tailored to their professional role. The responses to Q2 are reported comparatively in Table 7.5, while the responses to Q1 and Q3, together with the role-specific questions, are discussed in §7.3.2.

Participants provided written responses, which were analyzed qualitatively to extract both role-specific perspectives and recurring themes. The full Italian questions and responses are reported in Appendix H, along with English translations.

[10]Grammarly is a writing assistance tool that integrates with other applications to suggest spelling, grammar, and style improvements to users' text. See `https://www.grammarly.com/`.

Potential benefits	**Manager**	**Developer**	**Content Designer**
Time savings in content revision			
Increased stylistic uniformity		✓	✓
Reduced risk of publishing non-inclusive content	✓		✓
Training support for less experienced writers	✓		
Improved perceived quality of materials	✓		
Potential risks			
Workflow complications			
Stylistic flattening			
Undue or context-inappropriate rewrites going unnoticed	✓		✓
Errors due to inexperience with inclusive writing	✓		
Degraded perceived quality of materials	✓		

Table 7.5: Perceived benefits and risks of gender-inclusive writing support systems, as selected by participants in response to Q2. Checkmarks indicate the options selected by each participant.

7.3.2 Role-Specific Perspectives

The Manager's Perspective: Strategic Value and Organizational Change. The manager situates gender-inclusive writing within the broader context of corporate policies and standards. It is seen as particularly relevant for corporate content because, as the manager notes, inclusion already appears in human resources policies and sustainability certifications. From this perspective, equipping the organization with active writing support tools represents an opportunity to embed inclusive language practices "in the flow of work," facilitating change in corporate language that extends beyond training materials alone.

In response to Q1, the manager envisions a dual-function approach combining error signaling and rewriting suggestions. The manager's selections in Q2 (Table 7.5) reveal a measured awareness of trade-offs: while identifying potential benefits such as reduced risk of publishing non-inclusive content, formative support for less experienced writers, and improved perceived quality of materials, the manager simultaneously flags corresponding risks for each benefit. Inappropriate rewrites may go unrecognized, over-reliance on the tool may lead to errors, and perceived quality could degrade rather than improve. This pattern suggests an awareness that the tool's value depends critically on its implementation and on users' ability to engage with it critically rather than passively. This aspect is further discussed as a possible future direction in §8.4.3. Looking ahead (Q3), the manager expresses a desire to introduce inclusive language into work processes and, specifically for training courses, to integrate it into the company's certified social sustainability assurance system.

The manager's role-specific questions addressed client variability and technological control:

Q_M4. *Your clients may have different sensitivities regarding inclusive language. How should a rewriting tool manage this variability? Is it preferable to have uniform behavior, or to be able to calibrate the level of intervention based on the client or project?*

Q_M5. *This is certainly a topic for discussion; it would be necessary to share the rewriting model (e.g., prompting) with Diversity & Inclusion offices to accommodate their specific guidelines as well. The system will also need to evolve towards other possible sources of discrimination.*

The response to Q_M4 emphasized the need to share and align the rewriting model (including prompts) with Diversity and Inclusion offices, both internal and potentially those of clients, to accommodate specific guidelines. The manager also noted that the system should evolve to address other potential sources of discrimination beyond gender. On technological control, in response to Q_M5, the manager described the company's existing practice: API-based commercial systems like OpenAI's GPT[11] and Anthropic's Claude,[12] are used for creative projects without sensitive data, while local deployment is preferred when privacy is a concern. implying a trade-off between privacy and response quality, as further discussed in §8.3.3.

The Developer's Perspective: Trust, Explainability, and User Empowerment. The developer emphasizes a human-centered approach to system design, foregrounding user trust as a central concern. In response to Q1, the developer agrees with the dual-function model but adds a rationale absent from the other professional figures' responses: simple error flagging risks being unhelpful for users unfamiliar with inclusive writing guidelines, while direct intervention feels invasive and undermines control over the text. The developer's distinctive contribution is the emphasis on an explanatory layer: the system should not only highlight problematic passages and propose alternatives, but also *explain why* a formulation may be problematic. This design principle is further elaborated in §8.4.1.

The developer's selections in Q2 (Table 7.5) identify increased stylistic uniformity as the sole benefit, a practically-oriented consideration that aligns with the technical focus on system integration. No risks are selected, resulting in the most optimistic assessment among the three participants.

This explanatory component, further detailed in Q3, serves two complementary functions: building user trust by reducing the sense of having to blindly accept the model's suggestions, and providing educational value for less experienced writers. As the developer notes, "explaining certain edits serves both at the level of the user experience, reducing the feeling of having

[11] `https://openai.com/api/`
[12] `https://claude.com/platform/api`

to 'trust' the model, and at an educational level for less experienced content designers." With the addition of user-readable explanations, the tool functions not only as a support system for inclusive writing, but also as a training opportunity that the organization can leverage.

The developer's role-specific questions addressed technical integration and system customization:

Q_D4. *From a technical point of view, what characteristics are essential to integrate this tool in its ideal form into your workflow? Does the implementation add complexity? And how does this complexity weigh against the expected benefits?*

Q_D5. *How important is it to be able to modify the model (fine-tuning) and the prompts? Would you prefer a "closed" system that is easier to integrate, or a more flexible one that requires maintenance and specific skills in prompting and model training?*

On integration (Q_D4), the developer stresses embedding the tool within existing environments (e.g., as plugins for familiar interfaces) and ensuring the ability to handle concurrent workloads. For local deployment, compatibility with serving frameworks like vLLM [255] is identified as fundamental for enterprise adoption. Regarding customization (Q_D5), the developer expresses a clear preference for open, modifiable solutions: the ability to intervene on both prompts[13] and model weights is important because "inclusive writing can change depending on sector, audience, language, and so on." This variability calls for flexible systems capable of integrating multiple inclusive strategies, as discussed in §8.4.5. The long-term vision involves simplified interfaces that would allow even non-technical staff to modify prompts or initiate fine-tuning, democratizing control over the model: "a compromise that would allow everyone in the company to have a voice on the model to use and how to modify it based on the specific use case."

The Content Designer's Perspective: Workflow Integration and Professional Autonomy. The content designer provides detailed insights on how the tool should integrate into the creative writing process. In response to Q1, the system is envisioned as "a partner for open dialogue, capable of adding value to the writer's work without replacing their intention and style." This perspective positions the tool as fundamentally supportive rather than corrective, respecting writing as a creative and original process [171, 118].

The content designer articulates the same two-level operational model described by the manager and developer: a first level that highlights inclusivity issues through visual markers,

[13] In this regard, being able to prompt the model in the company's operating language, in this case Italian, becomes crucial.

and a second level that offers improvement proposals for the author to evaluate. This convergence across all three roles is significant: despite approaching the question from distinct professional concerns, all participants independently deemed fully automatic rewriting unsuitable, in favor of preserving the content designer's autonomy and final authority over the text [223, 171].

The content designer's selections in Q2 (Table 7.5) reflect a balanced assessment grounded in writing practice. Two benefits are identified, increased stylistic uniformity and reduced risk of publishing non-inclusive content, alongside one risk: undue or inappropriate rewrites going unnoticed. This combination reveals an appreciation for the tool's practical value while maintaining awareness that automated suggestions may not always align with the specific communicative context or authorial intent. The concern reinforces the content designer's emphasis on human oversight and final authority over textual decisions.

Looking ahead (Q3), the content designer emphasizes discretion as a guiding principle: the system should function as "a silent assistant that observes, signals, and suggests", comparable to driver-assistance systems that "offer support without assuming control," leaving the author "the creative responsibility and the final word."

The content designer's role-specific questions explored the tool's educational potential and pre-existing inclusive writing practices:

Q_C4. *Do you see this tool also as an opportunity to verify or improve your skills in inclusive writing?*

Q_C5. *Before this project, was the topic of inclusive writing taken into consideration in your work? Were there guidelines and tools, or was it left to individual sensitivity?*

About the tool's educational potential (Q_C4), the content designer offers a nuanced view: the value lies not in correction but in making visible elements that might escape notice during creative work, such as "formulations that are not inclusive, implicit stereotypes, linguistic choices that could be made more respectful or representative." Alternative suggestions become stimuli rather than solutions to accept automatically, offering opportunities "to expand one's repertoire and refine one's style." Notably, the response to the question about pre-existing practices (Q_C5) reveals that gender-inclusive language was already a structured concern before this collaboration: the content designer had personally drafted a document of best practices covering neutral language use, non-stereotyped examples and images, and consistency between text, tone, and audience. This context testifies to the growing interest in the industrial context for inclusive language and positions the tool as an enhancement to existing competencies rather than a corrective for deficiencies.

7.3.3 Synthesis, Implications, and Discussion

Despite approaching the question of gender-inclusive writing support from distinct professional concerns, the three participants converge on two fundamental points. First, they unanimously endorse a two-level interaction model: the system should signal potentially problematic passages and offer alternative formulations. Second, they unanimously advocate for preserving human agency, insisting that the content designer must retain final authority over the text. This consensus, emerging independently from organizational, technical, and creative perspectives, constitutes robust guidance for future system design. It also aligns with established principles in human-centered AI [19, 422], where research on human-in-the-loop systems has consistently shown that users prefer and perform better with systems that support rather than supplant their judgment [256, 171]. The content designer's driver-assistance metaphor captures this principle intuitively: the system should enhance human capabilities without assuming control. This is particularly important for complex tasks like inclusive writing, where suggestions carry normative weight and blind acceptance could undermine both the quality of outputs and the development of user competence [272].

The responses to Q2 (Table 7.5) reveal further patterns worth examining. No participant identifies workflow complications or stylistic flattening as risks, suggesting that integration concerns are minimal and that the tool is not perceived as disruptive or as a threat to expressive diversity. The benefits/risks pattern that emerges from the manager's responses, where for every benefit identified a corresponding risk is also flagged, reflects an awareness that outcomes depend on implementation quality and success in user engagement. The shared concern between the manager and the content designer about undue or context-inappropriate rewrites going unnoticed points to a critical challenge: the system must not only produce good suggestions but also support and educate users in recognizing when suggestions are inappropriate. This is where the developer's emphasis on explainability becomes crucial. The growing literature on interpretable NLP systems [380, 280, 109] supports this intuition: in the context of writing assistance, explainability serves not only to build trust but also to support learning, as users who understand why a formulation is flagged can internalize the underlying principles and apply them independently [37]. The educational potential of the tool is, indeed, recognized by all three roles. The generation of verbalized and user-readable explanation is a crucial future research direction, as discussed in §8.4.1.

The technical constraints identified by the manager and developer (local deployment, model size, and compatibility with standard and popular frameworks like vLLM) connect directly to our experimental work, and the connection reveals an interesting asymmetry between the two core components of a GNR support system. For classification, the results suggest that

small, dedicated models can effectively meet deployment requirements: the encoder-based classifier (0.11B parameters) outperformed all LLMs on the binary detection task while being orders of magnitude smaller and faster to run. For rewriting, the landscape is more varied. The dedicated model Inclusively (0.78B parameters), did not achieve competitive performance despite being specifically designed for this task (§7.2.2). The fine-tuning experiments (§7.2.3) on models ranging from 8B to 14B parameters showed improvements in neutralization rates, though with some trade-offs in meaning preservation, indicating that the generative complexity of rewriting benefits from larger and more capable models than classification. This pattern echoes the recognition-generation gap observed in cross-lingual settings (§6.2.3), where LLMs reliably identify when gender-neutrality is appropriate but struggle to consistently produce neutral translations. This generally confirms the dissociation between 'understanding' and execution, which represents a fundamental challenge for gender-inclusive language generation across both monolingual and cross-lingual tasks.

Among the tested systems, GPT-4.1 achieved the strongest performance on both neutrality and meaning preservation. As a commercial model requiring data to be sent to external servers, it may not suit all deployment scenarios, but it demonstrates that high-quality GNR is achievable with current state-of-the-art. For organizations prioritizing local deployment, the results above point to promising directions: the smaller fine-tuned open-weight models approached the performance of much larger systems, suggesting that progress in training data quality and curation could further narrow the gap. The analysis of metric-based data selection (§7.2.3) offers initial insights in this direction: the choice of what data to include in training directly shaped model behavior, with different filtering criteria producing different trade-offs between neutralization and meaning preservation. For a task like GNR, where success requires balancing competing objectives, developing principled data curation strategies will likely prove as consequential as architectural choices in determining real-world system performance [384].

More broadly, the investigation illustrates the gap between research prototypes and deployable systems in NLP: practitioners articulate requirements (plugin integration, concurrent workloads, local deployment, customization by non-technical staff) that are rarely central to academic evaluations [133]. Bridging this gap requires not only technical solutions but also the kind of stakeholder engagement documented here.

Finally, the content designer's confirmation that inclusive writing was already a structured concern before this collaboration carries an important implication for system design. The tool does not enter a vacuum but an environment with existing competencies, practices, and values. This observation aligns with participatory and value-sensitive approaches to AI development [55], which emphasize understanding user contexts rather than assuming that technology

fills an absence. For gender-inclusive writing support, this means designing systems that complement and enhance existing practices rather than positioning themselves as authoritative correctors.

Synthesizing these perspectives, four design principles emerge for gender-inclusive writing support systems in professional settings. While grounded in the monolingual GNR collaboration documented here, these principles extend naturally to cross-lingual GNT scenarios, where the need for human oversight is equally pronounced. Each principle emerges from the convergence of stakeholder insights and is informed by experimental findings from both monolingual and cross-lingual investigations.

1. **Transparency and explainability.** Users need to understand why a passage is flagged and why specific alternatives are proposed. This transparency builds trust, enables critical evaluation, and provides educational value. The explanatory layer should be integral to the system, not an afterthought.

2. **Seamless integration.** The tool should function within familiar editing environments rather than requiring context-switching. Technical architecture must support concurrent usage patterns typical of organizational settings, with compatibility with standard serving frameworks.

3. **Support (for) human judgment.** The system must implement a two-level interaction model (signal and suggest) and never apply automatic changes. Suggestions must be easy to accept, modify, or dismiss. The tool should enhance human capabilities without assuming control.

4. **Local deployment with customization.** Open-source models deployable on local infrastructure address privacy requirements and enable domain adaptation. The ability for both technical and non-technical staff to customize the system ensures it can evolve with organizational needs. Towards this direction, investment in training data curation may prove as consequential as architectural optimization in enabling smaller models to meet the quality requirements of professional deployment.

These principles extend beyond the specific collaboration documented here to inform the broader development of NLP tools for gender-inclusive language support, whether for monolingual rewriting or translation, aligning with established frameworks for human-centered AI design [380, 19, 422, 429]. More generally, this investigation illustrates how qualitative stakeholder engagement can complement quantitative evaluation, bridging the gap between technically successful systems and the practical and human needs of their intended users

[384, 272]. The convergence of perspectives from organizational, technical, and creative roles provides robust guidance that purely technical evaluations cannot offer, suggesting that similar multi-stakeholder investigations should accompany the development of other socially sensitive NLP applications.

Key Points

- **Acceptance of Automatic Rewriting**: All professional figures unanimously agree on the usefulness of the gender-inclusive writing support system.
- **Support over Replacement**: Gender-inclusive writing tools should function as partners that enhance human capabilities without assuming control. Beyond immediate assistance, they can serve as learning opportunities, helping professionals internalize inclusive writing principles and refine their skills over time.

  Two-Level System: Despite distinct professional concerns, organizational, technical, and creative roles independently converge on the preference for a two-level system signaling gendered sentences and suggesting gender-neutral reformulations.
- **Explainability for Trust and Learning**: Systems should explain in a user-readable way why formulations are flagged, to build trust by enabling critical evaluation of suggestions, and to provide educational value for less experienced writers.
- **Bridging Research and Practice**: Synthesizing stakeholder perspectives with experimental findings reveals that while high-quality GNR is technically achievable, successful deployment requires transparency, seamless workflow integration, support for human judgment, and local deployment with customization capabilities.

Chapter 8

Discussion and Conclusions

This thesis presented the research conducted throughout my PhD, which was dedicated to establishing gender-inclusive MT as a systematic area of investigation, addressing research questions spanning the conceptual, empirical, and practical dimensions of gender-inclusive MT, a domain that lacked systematic investigation despite its social relevance. Taken together, the work frames gender-inclusive MT as an end-to-end problem: defining when gender marking is warranted, translating those definitions into resources and evaluation procedures that support systematic measurement, and validating generation strategies under realistic constraints. The resulting perspective treats gender-inclusivity not as an isolated *bias correction* step, but as a research space shaped by linguistic variability, domain conventions, and user expectations. By connecting conceptual desiderata to empirical benchmarks and automated evaluation, the thesis provides a basis for comparing approaches across languages and inclusive strategies, and for diagnosing the current challenges to reliable generation. Finally, the work highlights that progress depends not only on model capability, but also on interaction and deployment choices, motivating designs that prioritize user control, transparency, and context-sensitive behavior.

This final Chapter first summarizes these contributions, then steps back to reflect on their boundaries, implications, and extensions. While the preceding Chapters demonstrated that gender-inclusive MT is relevant, feasible, and measurable, important questions remain about the conditions under which these methods can be responsibly deployed and how they might evolve as the field matures. This discussion situates the contributions within their current constraints, considers the ethical responsibilities accompanying systems designed to represent human identity, and identifies promising avenues for advancement.

The Chapter proceeds as follows. Section 8.1 summarizes the five primary contributions of the thesis spanning conceptual foundations, evaluation resources, evaluation methodologies,

generation approaches, and practical deployment. Section 8.2 discusses ethical dimensions including non-prescriptive design principles. Section 8.3 examines limitations related to data resources, evaluation subjectivity, model capabilities, and experimental scope. Section 8.4 outlines directions for future research, and Section 8.5 presents final comments on the broader significance of the work.

8.1 Summary of Contributions

This thesis makes five primary contributions spanning conceptual foundations, evaluation resources, evaluation methodologies, generation approaches, and practical deployment.

Conceptual Foundations (RQ1). Chapter 3 introduces a framework distinguishing *conservative* approaches to gender-inclusive MT, which leverage standard linguistic resources to avoid gendered forms, from *innovative* approaches employing neologistic devices such as neomorphemes. From a systematic analysis of institutional guidelines, the Chapter derives a taxonomy of neutralization strategies and articulates three desiderata specifying when and how neutralization should be applied. This formalization reframes the challenge from *de-biasing* (correcting erroneous gender assignments when cues exist) to *de-gendering* (avoiding gender marking unless gender information is explicitly supplied). This framing also motivates the thesis's scope: GNT constitutes the primary focus, since neutralization strategies operate within standardized grammar, enjoy broader institutional endorsement, and remain applicable across a wide range of communicative contexts. At the same time, neomorpheme-based translation is investigated as a complementary direction, recognizing that such forms address distinct representational needs and are gaining visibility, thereby extending the perspective from conservative strategies within existing norms to innovative solutions that expand the expressive possibilities of language.

Evaluation Resources (RQ2). Chapter 4 presents a suite of benchmarks that make gender-inclusive translation empirically tractable. GeNTE provides the first natural benchmark for English→Italian GNT, comprising 1,500 sentences with expert-crafted neutral references designed to test whether systems can neutralize appropriately while preserving gender when explicitly marked. The multilingual extension mGeNTE expands coverage to German, Spanish, and Greek through an annotation methodology balancing cross-linguistic consistency with language-specific conventions, enabling comparative investigation of how GNT capabilities vary across typologically diverse languages. Neo-GATE addresses neomorpheme-based translation through a flexible tagset mapping approach that accommodates multiple paradigms,

recognizing that no single neomorpheme system has achieved universal acceptance. Together, these resources transform gender-inclusive translation from abstract desiderata into a measurable research objective.

Evaluation Methods (RQ3). Chapter 5 demonstrates through a contrastive protocol that standard MT metrics fail to consistently reward correct neutralizations. This failure stems from high variability in valid neutral formulations and learned biases embedded in neural metrics trained on corpora reflecting masculine defaults. The Chapter develops two reference-free alternatives that overcome these limitations: a classifier-based method achieving strong performance for Italian, and an LLM-as-a-Judge framework that generalizes across Italian, German, Spanish, and Greek without dedicated training resources. The latter approach supports both sentence-level judgments and fine-grained phrase-level analyses, providing diagnostic capability beyond binary correctness labels. These methods enable scalable assessment of gender-inclusive translation quality without requiring expensive reference annotations for each new language or domain.

Generation Approaches (RQ4). Chapter 6 presents the first systematic empirical investigation of gender-inclusive translation generation. Initial experiments confirm that commercial MT systems and zero-shot LLMs fail to produce neutral outputs when translating gender-ambiguous content, defaulting instead to masculine forms. Few-shot prompting with a closed model (GPT-4) achieves approximately 65–70% neutral translations, demonstrating that LLMs possess latent capabilities that appropriate guidance can activate. However, manual evaluation reveals trade-offs between neutralization success and translation acceptability, with some neutral formulations introducing verbosity or awkwardness. The multilingual investigation using mGeNTE enabled systematic comparison across languages and models, revealing that performance varies substantially: German and Spanish yield higher neutral translation accuracy than Italian and Greek, likely reflecting differences in training data representation and the availability of gender-neutral linguistic resources in each language. A consistent finding across experimental settings is that closed commercial models substantially outperform open-weight alternatives, raising concerns about reproducibility, privacy, and long-term availability for deployment contexts involving sensitive content. The experiments also identify a recognition-generation gap, where models can often identify when neutrality is appropriate but struggle to produce corresponding neutral outputs. For neomorpheme-based translation, *mis-generation*, the inappropriate application of inclusive markers to elements that should not receive them, emerges as a distinctive error type requiring targeted mitigation strategies.

From Research to Practice (RQ5). Chapter 7 bridges academic research and practical deployment through collaboration with an Italian e-learning company, investigating gender-neutral rewriting as a monolingual task directly applicable to content revision workflows. Fine-tuned compact models match or exceed larger open-weight LLMs, with training data quality proving crucial. A key finding concerns the trade-off between optimizing for neutrality and meaning preservation: models trained on parallel gendered-neutral pairs filtered for high semantic similarity achieve better semantic fidelity but show reduced neutralization gains, highlighting the importance of balanced data curation strategies. Stakeholder investigation with professionals across organizational, technical, and creative roles reveals convergent preferences for a two-level interaction model where systems signal potentially problematic passages and offer suggestions but never apply automatic changes directly. This unanimous rejection of automatic rewriting not validated by humans yields four design principles for gender-inclusive writing support: transparency and explainability, seamless workflow integration, support for human judgment, and local deployment with customization capabilities. These principles position gender-inclusive language technology as a tool that empowers users rather than imposing particular linguistic choices.

8.2 Ethical Considerations

Research on gender-inclusive language technologies carries inherent ethical dimensions that extend beyond technical performance metrics. The systems and resources developed in this thesis aim to address representational harms arising from gender bias in MT (see §2.3). Yet the pursuit of inclusivity carries its own normative implications, particularly the need to balance advocacy for inclusive practices with respect for user autonomy and contextual variation.

A foundational principle underlying this thesis is that gender-inclusive translation represents one valuable approach to improving gender fairness in MT, not a universally superior solution to be imposed in all contexts. The desiderata articulated in §3.2 reflect this measured stance by limiting the scope of neutralization: gender should not be expressed when it cannot be properly assumed from the source, but when gender information is available through linguistic cues or external metadata, that information should be preserved.

The risks of over-prescriptive approaches are illustrated by the reception of automated inclusive writing features in commercial applications. When Google Docs introduced suggestions for gender-inclusive language [464], the system faced substantial user backlash, with critics objecting to suggestions they perceived as undue or overly broad [320]. This reception highlights a crucial lesson: well-intentioned interventions can backfire when they fail to

account for context, domain conventions, and user expectations.

The stakeholder investigation reported in §7.3 provides empirical grounding for non-prescriptive design. Despite approaching the question from distinct professional concerns, all three participants independently converged on the same functional specification: a two-level interaction model where systems signal potential issues and offer suggestions but never apply automatic changes. This unanimous rejection of automatic rewriting aligns with established principles in human-centered AI, where research has consistently shown that users prefer systems that support rather than supplant their judgment [19, 422, 256].

The distinction between conservative neutralization and innovative neomorpheme-based approaches (§3.1) further underscores the importance of non-prescription. Neither approach is inherently superior; rather, they address different needs and are appropriate in different contexts. Institutional communications may call for conservative strategies that align with established guidelines, while community spaces may embrace innovative forms that affirm non-binary visibility. Systems should therefore offer capabilities for multiple approaches rather than enforcing a single standard, allowing users and organizations to select strategies appropriate to their communicative contexts and values [152, 336].

Key Points

- **Non-Prescriptive Design**: Gender-inclusive translation represents an approach to improving gender fairness, not a universally superior solution to be imposed in all contexts. Systems should support user judgment rather than enforce particular linguistic choices.
- **Interaction Model**: Stakeholder consensus favors systems that signal potential issues and offer suggestions but never apply automatic changes, aligning with human-centered AI principles that emphasize user autonomy and control.
- **Context-Appropriate Strategies**: Neither conservative neutralization nor innovative neomorpheme-based approaches is inherently superior: they address different needs and are appropriate in different contexts. Systems should offer capabilities for multiple approaches rather than enforcing a single standard.

8.3 Limitations and Challenges

The contributions presented in this thesis are subject to constraints that simultaneously define the boundaries of the current work and point to broader challenges facing the field. Some of

these constraints reflect deliberate methodological choices, such as the focus on institutional text and sentence-level evaluation, that enabled controlled investigation but limit generalization. Others stem from the current state of resources and technology, including the scarcity of training data for gender-inclusive translation and the capability gap between closed and open models. Still others are intrinsic to the task itself: the inherent subjectivity of acceptability judgments and the tension between standardization and linguistic variation cannot be fully resolved through technical means alone. The following subsections examine these issues in detail, characterizing each as both a limitation of the present research and an ongoing challenge that future work must address.

8.3.1 Data Constraints

The evaluation resources developed in this thesis share a common foundation in institutional and parliamentary text, primarily drawn from the Europarl corpus [246]. This domain was deliberately chosen for its alignment with the formal communicative contexts where gender-neutral language guidelines are most established and where neutralization strategies are most readily applicable. The analysis of institutional guidelines in §3.1.1 revealed that such guidelines predominantly target formal, administrative, and professional communication. The survey reported in §4.1 confirmed that participants showed greater acceptance of neutral language in formal settings than in informal ones. Europarl thus represents a best-case scenario for GNT: the domain where inclusive language is both most expected and most feasible.

However, this focused domain coverage raises questions about generalization [162]. Translating text belonging to different domains and genres requires preservation of stylistic nuance and authorial voice [24, 237], where neutralization strategies might conflict with aesthetic goals. Moreover, technical documentation such as laws and contracts demands terminological precision [364], and neutralization approaches that introduce verbosity or alter established terminology could prove problematic in such contexts. Future work must investigate how approaches validated on institutional text transfer to these diverse contexts, and whether domain-specific adaptations or entirely different strategies become necessary.

A related limitation concerns the expert-crafted nature of the neutral references in the benchmarks. The neutralizations in GeNTE and mGeNTE were created by professional linguists with expertise in gender-inclusive language, following the institutional guidelines analyzed in §3.1.1. While this methodology ensures high linguistic quality and internal consistency, it may not capture the full range of acceptable neutralizations that native speakers would naturally produce. The COMMON-SET analysis reported in §4.2.3 quantifies this variability: when three professional translators independently neutralized the same sentences,

the majority exhibited substantial differences in neutralization strategies, confirming that multiple valid reformulations exist for any given gendered input. This finding connects to broader discussions in NLP about the limitations of gold-standard annotations and the value of capturing annotator disagreement as meaningful signal rather than noise [44, 349].

Beyond evaluation resources, the fine-tuning experiments in Chapter 7 illustrate the consequences of training data scarcity. While benchmark construction can proceed with carefully curated expert annotations, training dedicated systems requires substantially larger parallel corpora. The only available resource for Italian gender-neutral rewriting was a synthetic dataset generated through LLM-based paraphrasing rather than expert annotation (§5.2.1). As discussed in §7.2.3, this synthetic origin introduced quality concerns that necessitated filtering based on semantic similarity. The resulting trade-off between neutralization rate and meaning preservation underscores the need for large-scale, high-quality parallel corpora explicitly designed for gender-inclusive translation, a gap that constrains the development of dedicated systems and limits investigation of data scaling effects.

Finally, the thesis adopts a depth-first approach to language coverage, with Italian serving as the primary target language throughout. This methodological choice reflects the linguistic expertise required for work on gender-inclusive language: evaluating neutralization quality, crafting appropriate references, and assessing the acceptability of novel forms demands native-level proficiency and familiarity with ongoing sociolinguistic debates in each language community. Extending this depth of investigation to additional languages represents an important direction for future work, one that requires collaboration with researchers possessing the requisite linguistic and cultural expertise, as was the case for the extension of GeNTE into mGeNTE, which required collaboration with experts in Spanish, German, and Greek. More broadly, the challenge of creating evaluation resources that balance linguistic quality and scalability remains open: expert-crafted benchmarks ensure reliability but are expensive to produce, while crowdsourced or automatically generated alternatives risk introducing noise, biases, and trade-offs.

8.3.2 Evaluation Subjectivity

A fundamental challenge in evaluating gender-inclusive translation is that judgments of the acceptability of inclusive solutions are inherently subjective. Unlike tasks with clear-cut correctness criteria, gender-inclusive translation involves navigating preferences and contextual appropriateness that vary across individuals and communities. The survey conducted as part of GeNTE's development (§4.1) provides quantitative evidence of this variation: participant preferences were distributed across neutral translations, gendered alternatives, and judgments

of equivalence between the two. Preferences further varied by neutralization strategy and communicative context, with participants showing greater acceptance of neutral language in formal settings than in informal ones, and stronger endorsement of simple lexical substitutions compared to complex structural reformulations. These patterns reflect genuine variation in how speakers relate to gender-inclusive language rather than noise to be eliminated from evaluation [344, 387].

This subjectivity manifests concretely in annotation processes. The manual evaluation of GPT-generated translations reported in §6.1.3 yielded moderate inter-annotator agreement for acceptability judgments, reflecting the genuine difficulty of the task: evaluators legitimately assign different weights to the competing criteria of fluency, adequacy, and successful neutralization. Multiple neutralization approaches may be equally valid, and annotators' judgments are influenced by their individual exposure to and attitudes toward inclusive language forms [415]. As newly emerging linguistic practices, the perceived acceptability of neutral formulations is likely to evolve over time as speakers become familiar with these forms. Evaluation frameworks must therefore balance the need for reliable measurement with recognition that no single gold standard can capture the full spectrum of acceptable outputs [44, 150].

Such considerations extend to automatic evaluation as well. The classifier-based and LLM-as-a-Judge methods discussed in Chapter 5 provide scalable alternatives to manual annotation, but they cannot fully model the legitimate variation in human preferences. Automatic metrics cannot capture the trade-offs revealed by human evaluations: a neutralization that achieves perfect gender-neutrality but sounds stilted or verbose may be less acceptable overall than one that is slightly less neutral but maintains natural fluency. The inherent subjectivity of GNT quality assessment suggests that future systems should account for multiple valid solutions rather than targeting a single correct neutralization [352, 146, 349]. Users may have different preferences regarding trade-offs between preserving exact meaning, achieving complete neutrality, fluency, and acceptability, and systems might benefit from offering multiple neutralization options rather than a single output. Developing evaluation frameworks that can model these multidimensional trade-offs while respecting the inherent subjectivity of the task remains an open challenge. This challenge extends beyond gender-inclusive translation to broader questions in NLP evaluation: how to measure quality for tasks where multiple valid outputs exist, and how to incorporate legitimate variation in human judgment rather than treating it as noise to be averaged away.

8.3.3 Model Dependencies

The experimental work in this thesis reveals a persistent capability gap between closed commercial models and open-weight alternatives for gender-inclusive translation tasks. This gap manifests across multiple experimental settings and has significant implications for deployment.

The baseline experiments in §6.1 establish that commercial MT systems and zero-shot LLMs produce gender-neutral outputs for only a small fraction of gender-ambiguous source sentences, defaulting instead to masculine forms. Few-shot prompting with GPT-4 dramatically improves neutralization rates (§6.1.3), demonstrating that LLMs possess latent capabilities that appropriate prompting can activate. Yet the multilingual experiments in §6.2 reveal a persistent recognition-generation gap in open-weight models: they reliably identify when gender-neutrality is appropriate but fail to produce corresponding neutral outputs. This asymmetry indicates that successful neutralization requires more than task recognition: it demands proficiency in neutralization strategies in the target language and the ability to apply them while maintaining translation quality. Closed models currently offer superior capabilities for these tasks but raise concerns about reproducibility, privacy, and long-term availability (§2.1.4) [277, 503]. These concerns make such models unsuitable for many deployment contexts, particularly those involving sensitive content (§7.3).

Fine-tuning offers a path to improving open model performance, but the experiments in §7.2.3 reveal that success depends critically on training data quality. Models fine-tuned on synthetic gendered-neutral sentence pairs expose a fundamental trade-off: training on larger but noisier data achieves higher neutralization rates at the cost of reduced meaning preservation, while training on filtered high-quality data maintains semantic fidelity but shows reduced neutralization gains. This trade-off underscores the need for data curation strategies that balance both objectives and highlights the importance of human oversight in the rewriting process [288, 284, 424]. The stakeholder consensus reported in §7.3.3, which favored human-in-the-loop approaches over fully automatic rewriting, aligns with this empirical finding about the limitations of current automated systems. Such workflows can create a virtuous cycle in which practitioners progressively refine their inclusive writing skills while their interventions are collected as high-quality training data for open models. Over time, this synergy between human expertise and model improvement can reduce dependence on proprietary systems and enable fully open, locally controlled workflows. The underlying challenge, however, persists: developing open models that match closed alternatives for linguistically complex tasks requiring both broad world knowledge and fine-grained grammatical control. Until this gap narrows, practitioners face difficult trade-offs between capability, transparency, and

deployment constraints.

8.3.4 Scope Constraints

Perhaps the most significant constraint on the work presented in this thesis is its sentence-level scope. All benchmarks and experimental protocols operate on individual sentences presented in isolation, assessing each translation or rewrite independently without consideration of document context. This design choice reflects both the state of resources in the field and practical considerations for controlled experimentation, but it represents a significant simplification of real-world translation scenarios. Gender cues, role information, and communicative intent frequently unfold across multiple sentences and documents. A form that appears genuinely ambiguous in isolation may have its referent's gender established earlier in the document, while generically masculine formulations may recur in ways that only become visible when tracking patterns across an entire text. From this perspective, the sentence-level benchmarks introduced in Chapters 4 and 5 should be understood as controlled testbeds that approximate, but do not fully capture, the contextual complexity of actual translation scenarios.

A related, compounding simplification concerns the internal composition of the benchmark sentences. During corpus construction, sentences containing multiple human referents with different gender statuses were edited to retain only referents of the same type (§4.2.2), so that each entry presents a single, homogeneous gender phenomenon. While this ensures coherent sentence-level evaluation, it means that the benchmarks underrepresent configurations common in natural text, where a single sentence may simultaneously require neutralization for one referent and gender preservation for another, with annotation and generation decisions necessarily operating at the word or entity level. The results reported in Chapters 5 and 6 should therefore be understood as obtained under these controlled conditions, with performance on more complex inputs remaining an open question.

The sentence-level scope directly affects the applicability of the desiderata articulated in §3.2. Desideratum D1 specifies that gender should not be expressed when it cannot be properly assumed from the source, while Desideratum D2 requires preserving gender when reliably indicated through linguistic cues. These principles implicitly assume access to all relevant information for making the determination. In sentence-level evaluation, classification into categories requiring neutralization versus gendered translation is determined by examining each sentence in isolation. This classification is valid for independently evaluated sentences but may not hold when sentences appear in document context where prior discourse establishes referent gender. Without access to surrounding context, a sentence-level system cannot distinguish cases where neutralization is genuinely appropriate from cases where it would

override contextually established gender, potentially misgendering a referent whose gender was previously indicated.

A document-level view of GNT complicates these desiderata further by turning local gender expression decisions into discourse-level policies. Once a referent is introduced with a particular strategy, subsequent mentions should maintain consistency unless there is a principled reason to change. Document-level GNT therefore involves coordinating local neutralization decisions with global stylistic and pragmatic goals, and reconciling inclusive practices with constraints such as legal precision or organizational style guides. More broadly, document-level phenomena extend beyond coreference resolution to include tracking referent identity across sentence boundaries [299, 75], ensuring compatible gender expression for repeated mentions, and maintaining stylistic coherence throughout a document. Recent work has begun addressing these limitations, with benchmarks like GLITTER [356] introducing multi-sentence passages for evaluating gender-fair translation. However, such resources remain limited in language coverage and passage length.

Real translation tasks rarely involve isolated sentences: documents, articles, and institutional communications all involve extended discourse where sentences build upon one another [48, 519, 484]. The performance achieved on sentence-level benchmarks in this thesis should be interpreted as establishing capability on a simplified version of the task. Developing both resources and methods for document-level gender-inclusive translation represents a critical challenge for advancing the field. This challenge is not merely technical: it requires rethinking evaluation paradigms to assess discourse-level consistency, developing annotation methodologies that can capture document-wide gender expression policies, and designing systems capable of maintaining coherent strategies across extended text while remaining responsive to local contextual shifts.

Key Points

- **Domain Specificity**: The evaluation resources developed in this thesis draw from institutional and parliamentary text, representing a best-case scenario where gender-neutral language is both expected and feasible. Generalization to creative writing, technical documentation, and informal communication requires further investigation.
- **Evaluation Subjectivity**: Judgments of gender-inclusive translation acceptability are inherently variable across individuals and contexts. Evaluation frameworks must balance reliable measurement with recognition that no single gold standard can capture the full spectrum of perspectives.

- **Open vs Closed Models**: Commercial models currently outperform open-weight alternatives for gender-inclusive translation, but raise concerns about reproducibility, privacy, and long-term availability. Fine-tuning open models reveals trade-offs between neutralization rate and meaning preservation that depend critically on training data quality.
- **Sentence-Level Scope**: All benchmarks and experiments operate on isolated sentences, representing a significant simplification of real-world translation scenarios where gender cues, referents' identity, and communicative intent unfold across document context.

8.4 Future Research Directions

The contributions of this thesis define gender-inclusive MT as a tractable research area, but they also point to unresolved questions that shape a forward-looking agenda. The limitations discussed in §8.3 suggest that progress requires advances that extend beyond improving sentence-level neutralization accuracy: future work must address how inclusive behavior can be made transparent and controllable, how the inherent trade-offs between neutrality and other quality dimensions can be systematically understood and navigated, how approaches generalize across languages and domains, and how multiple inclusive strategies can be supported without collapsing linguistic and community variation into a single normative standard. The following directions outline concrete opportunities to build on the evaluation infrastructure and empirical findings developed in the preceding Chapters, with an emphasis on methods and resources that remain compatible with non-prescriptive, human-centered deployment.

8.4.1 Explainability

The stakeholder investigation reported in §7.3 identified explainability as a central requirement for gender-inclusive writing support systems. Industry practitioners emphasized that systems should not only flag potentially problematic passages and propose alternatives but also explain *why* a formulation may be problematic. This explanatory capacity serves dual functions: building user trust by enabling critical evaluation of suggestions, and providing educational value that helps users develop competence in inclusive writing over time [37]. The design principle of *transparency and explainability* synthesized in §7.3.3 positions this explanatory layer as integral to system design rather than a secondary addition.

Current work on explainability in this domain has focused primarily on post-hoc interpretability methods aimed at researchers rather than end-users. The context attribution analysis discussed in §6.2 employed feature attribution techniques to identify which prompt components influence model outputs, revealing that source category recognition and translation generation rely on different context signals. While such insights help explain phenomena like the recognition-generation gap and can inform prompt design, they do not directly translate into user-facing explanations. Bridging this gap between researcher-oriented interpretability and actionable user guidance represents a key direction for future research [131, 132].

Extending explainability methods to gender-inclusive translation requires addressing several domain-specific challenges. Explanations must operate at the linguistic level, mapping model behavior to meaningful grammatical categories that align with the neutralization strategies discussed in §3.1.2, rather than attributing importance to individual words or subwords. They should support learning rather than merely justifying system decisions, enabling users to build accurate mental models that foster independent competence [256]. Finally, the inherent subjectivity of gender-inclusive language (§8.3.2) requires explanations that acknowledge variability and contextual appropriateness rather than presenting suggestions as uniquely correct solutions, consistent with the non-prescriptive principles articulated in §8.2. Developing and evaluating explanatory interfaces that address these requirements constitutes a promising avenue for advancing gender-inclusive language technologies toward deployment.

8.4.2 Expanding Language Coverage

The evaluation resources developed in this thesis cover four target languages: Italian, Spanish, German, and Greek (§8.3.1). While these languages represent diverse grammatical gender systems, including two-gender Romance languages (Italian, Spanish), three-gender German, and Greek with its distinct script (§4.2.5), they constitute only a fraction of the world's grammatical gender languages. The cross-linguistic experiments reported in §6.2 revealed substantial variation in GNT performance across these languages, with German and Spanish yielding higher neutral translation accuracy than Italian and Greek. This variation likely reflects multiple interacting factors: differences in training data representation [230], varying availability of gender-neutral linguistic resources, and sociolinguistic factors such as the prominence of inclusive language discourse in different communities [487, 344]. Understanding how such factors influence model capabilities requires investigation across a broader range of languages.

Expansion should prioritize typological diversity. Languages with three-gender systems sometimes offer neutralization resources unavailable in Romance languages, though human referents typically remain marked [97]. Languages with different agreement patterns, such as

Hebrew and Arabic, pose distinct challenges not addressed by current resources [189, 365]. Lower-resource languages where inclusive practices may be less documented require particular attention, as the institutional guidelines analyzed in §3.1.1 predominantly reflect European and North American contexts. The methodology established through GeNTE and mGeNTE provides a template for such expansion (§4.2.1), though this requires collaboration with researchers possessing native-level proficiency and familiarity with sociolinguistic debates in each language community.

8.4.3 Interactive and User-Controlled GNT

The empirical findings in this thesis consistently highlight that there is no single notion of 'acceptable' gender-inclusive translation. The user survey in §4.1 and the manual evaluations in §6.1.3 both reveal substantial variation in preferences across strategies, domains, and individuals, as well as only moderate agreement on what counts as an adequate neutralization. This variability reflects genuine differences in how speakers relate to gender-inclusive language rather than annotation noise [415, 44]. Against this background, systems that commit to a single output risk imposing particular normative choices that may not align with user preferences or contextual requirements.

These observations motivate a shift from one-shot generation toward interactive, user-controlled workflows [416]. Instead of producing a single 'correct' neutralization, systems can present multiple alternatives that instantiate different strategies and trade-offs, allowing users to select or adapt the formulations that best fit their communicative goals and constraints. This perspective operationalizes the non-prescriptive stance articulated in §8.2: gender-inclusive translation is treated as a resource that empowers users, not as an authority that enforces uniform usage. The stakeholder investigation in §7.3 offers a concrete template for such interaction, converging on a two-level model where systems highlight potentially problematic passages and provide suggestions on demand, but refrain from unsupervised rewriting. This 'driver-assistance' view aligns with broader evidence from human-centered AI that users prefer systems that support rather than override their judgment [19, 422, 256].

Future work on interactive GNT opens several design and research questions. Preference specification mechanisms should allow configuration of neutralization behavior along dimensions such as pervasiveness, preferred strategies from the taxonomy in §3.1.2, and tolerance for verbosity, potentially at both individual and organizational levels [241]. The trade-off between neutralization and overall acceptability identified in this thesis (§6.1.3) constitutes a particularly important configurable dimension: some users may prioritize complete neutrality even at the cost of slight awkwardness, while others may favor natural-sounding formulations

that achieve sufficient rather than maximal neutralization. This dimension is discussed further in §8.4.4. Feedback signals from user choices can inform adaptive systems that learn over time which suggestions prove more useful in particular domains, creating the virtuous cycle discussed in §8.3.3 where human interventions simultaneously improve practice and generate high-quality training data. Finally, interface and explanation design become central: when presenting alternatives, systems must make differences and trade-offs intelligible without overwhelming users, building on insights from §8.4.1 and research on AI-assisted writing [171, 272]. Interactive, user-controlled GNT thus represents a natural extension of the thesis's emphasis on subjectivity, non-prescription, and human-in-the-loop workflows [429].

8.4.4 Understanding Trade-Offs in Gender-Neutral Translation

A recurring finding across the experimental work in this thesis is that achieving gender-neutrality often comes at a cost to other dimensions of output quality. The taxonomy of neutralization strategies presented in §3.1.2 anticipates this tension: while simple lexical substitutions may preserve fluency and brevity, they are not always applicable, and more complex reformulations can introduce verbosity, stylistic awkwardness, or semantic shifts. This trade-off manifests empirically throughout the thesis. The manual evaluation in §6.1.3 reveals that approximately 20–30% of neutral outputs fall into borderline acceptability categories, with evaluators citing concerns about fluency or adequacy. The fine-tuning experiments in §7.2.3 expose a fundamental tension between optimizing for neutralization rate and preserving semantic fidelity. These patterns indicate that the trade-off reflects an inherent characteristic of gender-inclusive language generation rather than a byproduct of current approaches.

The trade-off operates along multiple dimensions. Verbosity represents perhaps the most visible cost: periphrastic neutralizations such as replacing *giornalisti*$_{[M]}$ (EN: journalists) with *persone che lavorano nel giornalismo* (EN: people working in journalism) achieve neutrality but substantially increase length and may reduce readability [106]. Semantic precision constitutes a second dimension: the example above shifts meaning, since 'people working in journalism' are not necessarily journalists. Fluency and naturalness form a third dimension, as neutralized formulations may sound awkward even when technically correct. Finally, stylistic consistency matters in domains with established conventions, where neutralization strategies may clash with register expectations or authorial voice [24, 237]. These dimensions are not independent: a reformulation that reduces verbosity may compromise semantic precision, while one that maintains naturalness may require accepting incomplete neutralization.

The inherent subjectivity discussed in §8.3.2 compounds the challenge of modeling these trade-offs, as the appropriate balance depends on user preferences, domain requirements,

and communicative goals [241, 435]. Future evaluation frameworks must therefore move beyond binary neutrality assessment toward multi-dimensional models that can capture these trade-offs explicitly [279, 385, 50]. Such research would directly inform the interactive systems discussed in §8.4.3: rather than presenting users with a single output, systems could offer alternatives along the trade-off, allowing users to select formulations matching their priorities. More broadly, this direction connects gender-inclusive translation to fundamental questions in natural language generation about balancing competing objectives and designing systems that navigate multi-dimensional quality spaces in accordance with user preferences [325, 213, 170, 509].

8.4.5 Integrating Multiple Inclusive Strategies

This thesis investigated conservative neutralization and neomorpheme-based approaches as largely separate tracks, reflecting both the distinct challenges they pose and the different contexts in which they are appropriate. Conservative strategies align with standardized grammar and enjoy broad institutional endorsement, making them suitable for formal and professional communication. Neomorphemes, by contrast, emerge from grassroots efforts to explicitly represent non-binary identities and remain largely confined to informal contexts and specific communities. Neither approach is universally superior; rather, they address different representational needs and carry different trade-offs regarding accessibility, acceptance, and expressive power (§3.1).

Future gender-inclusive translation systems should move beyond this separation toward integrated approaches capable of deploying multiple strategies as context requires. Such systems would need mechanisms for selecting appropriate strategies based on several factors: the referent's established preferences when known, the formality and domain of the text, the intended audience, and accessibility requirements. A legal document might call for conservative neutralization to maintain institutional acceptability, while a community newsletter might appropriately employ neomorphemes that affirm non-binary visibility. When translating content that discusses a specific individual, the system should respect that person's expressed preferences for linguistic reference whenever such preferences are known or can be inferred.

The evaluation infrastructure developed in this thesis provides foundations for such integration. Neo-GATE's flexible placeholder annotation system (§4.3.2) can accommodate any neomorpheme paradigm through tag mapping, enabling evaluation across multiple inclusive strategies within a unified framework. Recent work has begun exploring resources that combine conservative and innovative forms, pointing toward evaluation approaches that can assess systems capable of both [356]. Developing generation methods and evaluation

frameworks that handle the full spectrum of gender-inclusive strategies, rather than treating them as isolated tasks, represents an important direction for creating translation systems that can serve diverse users and contexts with appropriate linguistic choices.

8.4.6 Improved Training Methods and Data

A fundamental obstacle underlying many challenges identified in this thesis is the scarcity of training data for gender-inclusive translation. Large-scale parallel corpora featuring consistent GNTs do not exist: standard MT training data was not created or collected with inclusivity as a consideration, and naturally occurring translations overwhelmingly employ default gendered forms [475, 405, 112]. This absence shapes model behavior at a foundational level, as both encoder-decoder systems and LLMs learn from data that systematically underrepresents the very patterns they would need to produce inclusive outputs.

The synthetic data generation approach developed for classifier training (§5.2.1) demonstrates both the necessity and the limitations of current workarounds. While LLM-generated data can fill immediate gaps, repurposing such data for generative fine-tuning reveals quality concerns: some generated pairs exhibit semantic divergence that, while acceptable for classification, proves problematic when the goal is to learn faithful reformulation. The fine-tuning experiments in §7.2.3 exposed a fundamental trade-off between neutralization and meaning preservation that stems directly from training data characteristics. These findings suggest that principled data curation strategies may prove as consequential as architectural innovations in determining system performance [384].

Future progress requires moving beyond synthetic data toward high-quality parallel corpora explicitly designed for gender-inclusive translation. Such resources should be created with expert annotation that balances competing objectives: pairs must differ in gender expression while remaining semantically equivalent, and the range of valid neutralization strategies should be represented without biasing toward any single approach. The recognition-generation gap identified in §6.2 suggests that training data should specifically target the generation of neutral formulations, providing extensive examples of well-formed neutral outputs across diverse neutralization strategies.

The interactive systems discussed in §8.4.3 offer a promising pathway for data collection. Deployed systems that support human-in-the-loop workflows can progressively accumulate expert decisions as users accept, modify, or reject suggestions [494]. This creates potential for a virtuous cycle: practitioners refine their inclusive writing skills while their interventions are collected as high-quality training signal, enabling model improvements that reduce dependence on synthetic data over time. Developing infrastructure for such collection, along with

methodologies for balancing the multiple objectives inherent in gender-inclusive translation, represents an essential foundation for advancing the field beyond current limitations.

Key Points

- **Explainability**: Future systems should not only flag potentially problematic passages but also explain why a formulation may be problematic, building user trust and providing educational value that helps users develop competence in inclusive writing over time.
- **Language Coverage Expansion**: Current resources cover four target languages (Italian, Spanish, German, Greek), representing only a fraction of grammatical gender languages. Expansion should prioritize typological diversity, but requires collaboration with researchers possessing native-level proficiency in each language community.
- **Interactive and User-Controlled Systems**: Instead of committing to a single output, future systems could present multiple alternatives instantiating different strategies and trade-offs, allowing users to select formulations that best fit their communicative goals and constraints.
- **Neutrality-Acceptability Trade-off**: Gender-inclusive translation involves inherent tension between achieving complete neutrality and preserving readability, fluency, and semantic fidelity. Future work should develop frameworks for systematically characterizing these trade-offs across different strategies and evaluation methods that capture multi-dimensional quality.
- **Integration of Multiple Strategies**: Future systems should move beyond treating conservative neutralization and neomorpheme-based approaches as separate tracks toward integrated approaches capable of dynamically deploying multiple strategies as context demands.
- **Improved Training Data**: Progress requires moving beyond synthetic data toward high-quality parallel corpora explicitly designed for gender-inclusive translation, potentially leveraging human-in-the-loop workflows that simultaneously improve practice and generate training signal.

8.5 Concluding Remarks

This Chapter has situated the contributions of the thesis within a broader context of limitations, ethical considerations, and future possibilities to cover remaining gaps and for impactful research. The discussion reveals gender-inclusive MT as a research area characterized by intrinsic complexity: technical challenges interweave with sociolinguistic questions, and progress requires attention to both.

Taken together, the limitations, ethical issues, and future directions discussed here point to a long-term research agenda rather than a problem that can be solved by a single model or metric. The subjectivity of acceptability, the context-dependence of appropriate strategies, and the need to reason beyond isolated sentences all suggest evaluation frameworks that move beyond reference matching toward user- and community-centered notions of quality.

Ultimately, the goal of gender-inclusive MT is not to impose a particular linguistic ideology but to expand the expressive possibilities available to both human users and automatic systems. Current MT technology constrains these possibilities, defaulting to masculine forms that may misrepresent, exclude, or harm the individuals and communities that translation is meant to serve. The work presented in this thesis takes steps toward removing those constraints, providing resources that enable evaluation of inclusive capabilities, methods that assess whether systems achieve inclusivity, and approaches that can generate inclusive outputs when appropriate.

In this light, the enduring legacy of this thesis for those who will continue and advance this research lies in having demonstrated that gender-inclusive machine translation can be pursued in a structured, replicable way: by clearly identifying what constitutes gender-inclusive behavior, by making explicit the trade-offs between different strategies, and by designing tools that can be adapted to new languages, domains, and stakeholders as they enter the conversation.

Appendix A

Gender-Inclusive Language Guidelines

The following guidelines for gender-inclusive language were analyzed:

E1 United Nations Economic Commission for Western Asia (2014)

E2 United Nations (2018)

E3 General Secretariat, Council of the European Union (2018)

E4 European Parliament (2018)

E5 North Atlantic Treaty Organization (2020)

E6 Australian Government (2021)

E7 University of Houston (2022)

E8 Australian National University (n.a.)

E9 United Nations Women (n.a.)

E10 University of North Carolina at Chapel Hill (n.a.)

E11 University of Pittsburgh (n.a.)

E12 Royal Melbourne Institute of Technology (n.a.)

E13 California State University San Marcos (n.a.)

E14 University of Otago (n.a.)

E15 The University of Texas at Austin (n.a.)

I1 Cancelleria Federale Svizzera (2012)

I2 Università di Torino (2015)

I3 Università degli Studi di Padova (2017)

I4 Segretariato Generale, Consiglio dell'Unione Europea (2018)

I5 Parlamento Europeo (2018)

I6 Università degli Studi di Verona (2020)

I7 Università di Bologna (2020)

I8 Università degli Studi dell'Aquila (2020)

I9 Università di Siena (2021)

I10 Istituto Universitario Federale per la Formazione Professionale (2021)

I11 Università della Calabria (2021)

I12 Università degli Studi di Milano (2021)

I13 Università Mediterranea di Reggio Calabria (n.a.)

I14 Università di Trento (n.a.)

I15 Università di Ferrara (n.a.)

Appendix B

GeNTE Corpus Details

This appendix reports corpus-level details that complement the discussion in Chapter 4. It documents the editing interventions applied during the construction of GeNTE (§B.1), summarizes recurrent difficulties encountered when producing gender-neutral references (§B.2), and quantifies the degree of linguistic diversity across independently produced references in the COMMON-SET (§B.3). The material is intended to support interpretation and reproducibility, and is not required for following the main narrative.

B.1 Data Editing Report

Two types of interventions are applied during corpus construction. A first set of *functional* interventions supports corpus usability and experimental control, in line with the design principles described in §4.2.1. A second set of *quality-oriented* interventions improves cleanliness and lexical diversity by correcting errors, removing artifacts, and reducing unintended repetition.

Functional interventions include two procedures. First, source and reference sentences are edited to ensure that all human referents within an entry require the same type of realization in the target language (either consistently neutral or consistently gendered and expressing the same gender). This procedure affects 203 entries. Second, gendered entries are duplicated to create minimal pairs differing only in the grammatical gender of the relevant terms: gender-marked words are replaced with their opposite-gender equivalents, yielding 126 masculine and 247 feminine entries (373 total). Since some entries undergo both procedures, functional interventions affect 576 entries overall.

Quality-oriented interventions correct translation issues in the original references and remove extraneous elements in both sources and references. For instance, the segment “EN)”

is removed from the source sentence "EN) I would like, in particular, to thank Mrs Van den Burg, a Dutch Social Democrat who worked particularly hard on Article 25." Such corrections are applied to 89 entries. In addition, lexical repetition is reduced by replacing the most frequent noun that systematically triggers gender marking, *rapporteur*, with alternative terms from the institutional and administrative domain (e.g., *spokesperson*, *delegate*, *deputy*). This operation is performed on 70 entries. Overall, 314 original source sentences and 393 original reference sentences are edited.

B.2 Challenges in the Creation of Gender-Neutral References

The creation of gender-neutral references involves systematic challenges that arise from the interaction between Italian morphology and syntax and the operational goal of producing fluent translations that preserve the propositional content of the gendered reference while avoiding unnecessary gender marking. Two recurring phenomena are observed during reference creation.

- **Articles:** In 11 instances, translators produce partial neutralizations because masculine articles remain in the target sentence. This suggests that, even for native speakers, articles may be less salient cues of gender than nouns, adjectives, or participles. All cases are identified during linguistic validation and corrected.

- **Lexical gender:** In 4 instances, translators are unable to neutralize lexically gendered nouns such as *sorella* (EN: sister) and *figlia* (EN: daughter). These cases occur in the construction of neutral references for SET-G, which supports contrastive evaluation (§5.1) and therefore requires neutral strategies even for unequivocally gendered terms. Resolving such cases typically entails rephrasing that departs more substantially from the surface form of the gendered reference.

Additional difficulties arise less systematically for domain-entrenched terms, such as *deputato* (deputy), where conventional translations are strongly lexicalized and gender-marked. In these cases, neutralization may be perceived as counter-intuitive and may require periphrastic solutions (e.g., *persona deputata*, EN: deputed person). All such cases are resolved through intervention by the linguist.

B.3 Linguistic Diversity in Gender-Neutral References

Table B.1 quantifies linguistic diversity among the COMMON-SET references by reporting pairwise BLEU scores computed by matching each reference of an entry against the other references for the same entry. The scores indicate that the independently produced references remain relatively close in surface form, which is expected given that they preserve the same underlying content and differ primarily in gender-related realizations. At the same time, the fact that scores remain below 100 BLEU points confirms that the references are not identical and reflect alternative neutralization choices.

COMMON-SET-G

↓ REF CAND →	Reference 1	Reference 2	Reference 3	REF-G
Reference 1	-	75.14	77.65	74.14
Reference 2	75.14	-	75.09	72.08
Reference 3	77.59	75.03	-	74.89
REF-G	74.04	71.98	74.82	-

COMMON-SET-N

↓ REF CAND →	Reference 1	Reference 2	Reference 3	REF-G
Reference 1	-	76.88	76.27	75.89
Reference 2	76.91	-	76.15	73.36
Reference 3	76.28	76.14	-	73.02
REF-G	75.78	73.26	72.92	-

Table B.1: BLEU scores representing the linguistic variability in COMMON-SET's references.

Diversity is consistent across sets. In COMMON-SET-N, pairwise BLEU scores among neutral references are tightly clustered, with differences below one BLEU point between the highest and lowest values, suggesting that alternative strategies have comparable impact on surface similarity. This characterization provides context for the reference-based evaluation results discussed in Chapter 5: even when content is held constant, neutralization can induce measurable variation, motivating the use of multiple references and caution in interpreting surface-similarity scores as proxies for neutrality.

Appendix C

Neo-GATE Corpus Details

C.1 Tagset and Annotation

Table C.1 reports the complete tagset used in NEO-GATE, as well as the tagset mappings for the Schwa and the Asterisk paradigms.

TAG	Description	Masculine	Feminine	Asterisk	Schwa
<ENDS>	inflectional morpheme (word ending), singular	o, e, tore	a, essa, trice	*	ə
<ENDP>	inflectional morpheme (word ending), plural	i, tori	e, esse, trici	*	ɜ
<DARTS>	definite article, singular	il, lo, l'	la, l'	l*	lə
<DARTP>	definite article, plural	i, gli	le	l*	lɜ
<IART>	indefinite article	uno, un	una, un'	un*	unə
<PARTP>	partitive article, plural	dei, degli	delle	de*	deɜ
<PREPdiS>	articulated preposition with root 'di', singular	del, dello, dell'	della, dell'	dell*	dellə
<PREPdiP>	articulated preposition with root 'di', plural	dei, degli	delle	dell*	dellɜ
<PREPaS>	articulated preposition with root 'a', singular	al, allo, all'	alla, all'	all*	allə
<PREPaP>	articulated preposition with root 'a', plural	agli, ai	alle	all*	allɜ
<PREPdaS>	articulated preposition with root 'da', singular	dal, dallo, dall'	dalla, dall'	dall*	dallə
<PREPdaP>	articulated preposition with root 'da', plural	dagli	dalle	dall*	dallɜ
<PREPinP>	articulated preposition with root 'in', plural	negli	nelle	nell*	nellɜ
<PREPsuS>	articulated preposition with root 'su', singular	sul, sullo, sull'	sulla, sull'	sull*	sullə
<PREPsuP>	articulated preposition with root 'su', plural	sugli	sulle	sull*	sullɜ
<DADJquelS>	demonstrative adjective (far), singular	quel, quello, quell'	quella, quell'	quell*	quellə
<DADJquelP>	demonstrative adjective (far), plural	quegli	quelle	quell*	quellɜ
<DADJquestS>	demonstrative adjective (near), singular	questo, quest'	questa, quest'	quest*	questə
<DADJquestP>	demonstrative adjective (near), plural	questi	queste	quest*	questɜ
<POSS1S>	possessive adjective, 1st person singular, singular	mio	mia	mi*	miə
<POSS1P>	possessive adjective, 1st person singular, plural	miei	mie	mi*	miɜ
<POSS2S>	possessive adjective, 2nd person singular, singular	tuo	tua	tu*	tuə
<POSS2P>	possessive adjective, 2nd person singular, plural	tuoi	tue	tu*	tuɜ
<POSS3S>	possessive adjective, 3rd person singular, singular	suo	sua	su*	suə
<POSS3P>	possessive adjective, 3rd person singular, plural	suoi	sue	su*	suɜ
<POSS4S>	possessive adjective, 1st person plural, singular	nostro	nostra	nostr*	nostrə
<POSS4P>	possessive adjective, 1st person plural, plural	nostri	nostre	nostr*	nostrɜ
<PRONDOBJS>	direct object pronoun, singular	lo	la	l*	lə
<PRONDOBJP>	direct object pronoun, plural	li	le	l*	lɜ

Table C.1: The full tagset used in NEO-GATE and the tagset mappings to the Italian gendered forms and the desired forms in the Asterisk and Schwa nomorpheme paradigms.

Appendix D

Prompts and Exemplars

This appendix collects prompt templates and in-context exemplars used throughout the experimental chapters, besides the ones reported in the main chapters. Prompts are reported in a model-agnostic form (typically as system-role instructions), and are instantiated by inserting task-specific inputs in placeholders (e.g., <{input sentence}>). The section structure mirrors the thesis', from synthetic data generation and GNT prompting, to translation-quality checks, LLM-as-a-Judge evaluation, and gender-neutral rewriting.

D.1 Synthetic Data Generation Prompts

Tables D.1 and D.2 report the two-stage prompting pipeline used to construct the synthetic Italian dataset described in §5.2.1. The first-round template generates controlled sentence triplets that differ only in the inserted seed term (neutral, masculine, feminine), yielding tightly aligned minimal pairs. The second-round template rewrites these triplets into longer, Europarl-style variants while enforcing the same gender constraints, increasing syntactic and contextual diversity for training and validation of the reference-free classifier in §5.2.

D.2 GNT Prompts

Table D.3 lists the full set of demonstration triplets used for in-context learning in the GNT prompting experiments presented in §6.1.2. Exemplars are organized into the *seen* (S) and *not seen* (NS) conditions introduced in §6.1, which support analysis of whether prompt-based neutralization generalizes to gendered terms that do not appear in the demonstrations. For readability, terms that drive the gendered versus neutral contrast are highlighted in bold, and glosses are provided in square brackets for the neutral reformulations.

Generate Italian sentences starting from the triplet of `seed words`.

Generate the sentences based on these instructions:
- Output 25 groups of three sentences, one sentence per each seed word of the triplet.
- The three sentences in each group must differ just in the seed word.
- Use one 'seed word' at a time in each sentence.

Expected output example:

1.
- A seguito degli incontri parlamentari, le commissioni hanno deciso di perseguire stabilità internazionale.
- A seguito degli incontri parlamentari, i commissari hanno deciso di perseguire stabilità internazionale.
- A seguito degli incontri parlamentari, le commissarie hanno deciso di perseguire stabilità interazionale.

2.
- Se le commissioni non si decidono, dobbiamo riunirle.
- Se i commissari non si decidono, dobbiamo riunirli.
- Se le commissarie non si decidono, dobbiamo runirle.

...

25.
...

Table D.1: Prompt template for the generation of triplet of sentences from (NEUT/FEM/MASC) seed words.

Rewrite each `seed` of Italian sentences based on the following instructions:

a) Change the structure of the sentences, make them longer, and with a style
b) Make the sentence longer.
c) Use a style that resembles the language from the European Parliament.

Also, the rewritten sentences must adhere to the following constraints:
i) the rewritten `Neutral seed` does not contain masculine or feminine occupational nouns.
il) the rewritten `Masculine seed` must contain masculine occupational nouns, but does not contain feminine occupational nouns.
iii) the rewritten `Feminine seed` must contain feminine occupational nouns, but does not contain masculine occupational nouns.

Think step by step.
According to the instructions and constrain output 6 groups of three sentences per each triplet of `seed`.

Expected output example:

1.
- In seguito ai colloqui parlamentari, le commissioni hanno deliberato di perseguire la stabilità internazionale e della cittadinanza.
- In seguito ai colloqui parlamentari, i commissari hanno deliberato di perseguire la stabilità intemazionale e dei cittadini.
- In seguito ai colloqui parlamentari, le commissarie hanno deliberato di perseguire la stabilita internazionale e delle cittadine.

2.
- Nel caso in cui le commissioni non pervengano a una decisione, dovremmo procedere alla loro riunione e convocare una seconda udienza come stabilito dall'onorevole.
- Nel caso in cui i commissari non pervengano a una decisione, dovremmo procedere alla loro riunione e convocare una seconda udienza come stabilito dal signor onorevole.
- Nel caso in cui le commissarie non pervengano a una decisione, dovremmo procedere alla loro riunione e convocare una seconda udienza come stabilito dalla signora onorevole. ...

6.
...

Table D.2: Prompt template for the rewriting of triplet of (NEUT/FEM/MASC) seed sentences.

D.3 Translation Prompt

Table D.4 reports the three-shot translation prompt used to elicit plain English→Italian translations in a fixed interaction format. This template supports the baseline translation-quality verification described in §6.3.1 and the preliminary MT checks in §5.3. The consistent dialogue structure, together with explicit delimiters, facilitates automatic extraction of the translation output and reduces post-processing issues caused by additional model commentary.

D.4 LLM-as-a-Judge Prompts

Tables D.5–D.8 report the system role messages for the LLM-as-a-Judge prompt families introduced in §5.3.1. The MONO variants operationalize target-only evaluation, whereas the CROSS variants condition evaluation on both the source and the target to assess gender correctness in translation, as described in §5.3. For each mode, the `L` prompts request a sentence-level decision only, while the `P+L` prompts require phrase-level extraction and labeling before the final decision, supporting the analysis in §5.3.2 of how intermediate annotations affect reliability across languages and models.

D.5 GNR Prompts

Table D.9 reports the system messages used in the few-shot prompting experiments for gender-neutral rewriting described in §7.2.2. Two instruction formats are contrasted: a concise directive (GFG) and a longer specification (REWRITE) that enumerates admissible rewriting strategies and constraints, each provided in both Italian and English. Table D.10 reports the prompts used in the automatic classification experiments in §7.2.4, including a binary setup (neutral versus gender-marked) and a ternary setup that isolates inputs that contain no references to human beings.

	Seen
SRC	Secondly, how far does it increase transparency and accountability **of the MEPs**?
GEND	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **dei parlamentari europei**?
NEUT	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **dei membri del Parlamento Europeo** [of the members of the European Parliament]?
SRC	**President**, **everyone** must continue to adopt an ambitious approach on these issues.
GEND	**Signora Presidente**, su tali questioni sarà necessario che **tutti** continuino a dare prova d'ambizione.
NEUT	**Presidente** [President], su tali questioni sarà necessario che **ogni persona** [every person] continui a dare prova d'ambizione.
SRC	**Several fishermen** have **joined** with **the politicians** in Belgrade.
GEND	A Belgrado, **molti pescatori** si sono **schierati** dalla parte **dei politici**.
NEUT	A Belgrado, **molte persone che lavorano nella pesca** [many people who work in fishery] hanno **preso le parti** [have taken the side of] di **chi fa politica** [of those who engage in politics].
	Not seen
SRC	Secondly, how far does it increase transparency and accountability **of the writers**?
GEND	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**?
NEUT	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive** [of those who write]?
SRC	**HR manager**, **the employees** must continue to adopt an ambitious approach on these issues.
GEND	**Direttore delle risorse umane**, su tali questioni sarà necessario che **gli impiegati** continuino a dare prova d'ambizione.
NEUT	**Responsabile delle risorse umane** [HR manager], su tali questioni sarà necessario che **il personale** [the staff] continui a dare prova d'ambizione.
SRC	**Several freshmen** have **joined** with **the musicians** in Belgrade.
GEND	A Belgrado, **molti studenti del primo anno** si sono **schierati** dalla parte **dei musicisti**.
NEUT	A Belgrado, **molte matricole** [many first-years] hanno **preso le parti** [have taken the side of] **delle persone del mondo della musica** [of the people in the music business].

Table D.3: All the <source sentence, gendered translations, and neutral translations> triplets used as demonstrations in both the S and NS sets of examples. Relevant terms for the gendered/neutral comparison are in bold. GNT glosses are available in square brackets.

Role	Content
`user`	Translate the following English sentence into Italian: <Construction is ongoing for five new skyscrapers at the site, with a transportation center and memorial park in the middle.>
`assistant`	<Nel sito sono in corso i lavori di costruzione di cinque nuovi grattacieli, con un centro trasporti e un parco alla memoria nel mezzo.>
`user`	<The game is based on the Second Battle of Fallujah, a vicious battle between American and Iraqi forces.>
`assistant`	<Il videogioco simula la seconda battaglia di Fallujah, un conflitto cruento tra le forze armate americane e quelle irachene.>
`user`	<Born in the Croatian capital, Zagreb, Bobek gained fame while playing for Partizan Belgrade.>
`assistant`	<Nato nella capitale croata Zagabria, Bobek è diventato famoso giocando nel Partizan Belgrado.>
`user`	<`{input sentence}`>

Table D.4: The 3 shots prompt used in the general translation preliminary experiments.

You are a language expert specializing in evaluating gender neutrality in Italian texts. Your task is to assess each provided sentence and determine whether it is gendered or neutral. Guidelines: 1. Identify relevant phrases: carefully analyze the Italian sentence and focus on all phrases that refer to human beings or groups of human beings, including: - Noun phrases (e.g., "un'ottima oratrice", "la cittadinanza"), - Verb phrases (e.g., "è molto felice", "ho purtroppo dovuto"), - Adjective phrases (e.g., "felicemente sposato", "molto competente"). 2. Evaluate gender information: consider only the social gender conveyed by the phrases, not grammatical gender. For example: - Phrases like "un oratore", "è molto contento", "tutti i colleghi", and "i cittadini" are masculine; - Phrases like "un'oratrice", "è molto contenta", "tutte le colleghe", and "le cittadine" are feminine; - Phrases like "una persona che parla in pubblico", "è molto felice", "tutte le persone con cui lavoro", and "la cittadinanza" do not express social gender, therefore they must be considered neutral. 3. Assign a label: - If all references to human beings are gender-neutral, label the sentence as "NEUTRAL". - If one or more expressions convey a specific masculine or feminine gender, label the sentence as "GENDERED".

Table D.5: System message for prompt MONO-L (Italian).

You are a language expert specializing in evaluating gender neutrality in German texts. Your task is to extract target German phrases that refer to human beings and determine whether each phrase is masculine, feminine, or neutral. Based on the phrases, assess whether the sentence is gendered or neutral.

Guidelines:

1. Identify relevant phrases: carefully analyze the German sentence and focus on all phrases that refer to human beings or groups of human beings (e.g., "eine ausgezeichnete Rednerin", "die Bürgerschaft", "Sie").

2. Evaluate gender information: consider only the social gender conveyed by the phrases, not grammatical gender, and assign a label to each phrase [M/F/N]. For example:
- Phrases like "Ein Redner", "Der Student", "Der Bürger", and "alle Kollegen" are masculine [M];
- Phrases like "Eine Rednerin", "Die Studentin", "Die Bürgerinnen", and "alle Kolleginnen" are feminine [F];
- Phrases like "Eine referierende Person", "Die Studierenden", "Die Bürgerschaft", and "alle Kollegiumsmitgliedern" do not express social gender, therefore they must be considered neutral [N].

3. Assign a sentence-level label:
- If all references to human beings are gender-neutral, label the sentence as "NEUTRAL".
- If one or more phrases convey a specific masculine or feminine gender, label the sentence as "GENDERED".

Table D.6: System message for prompt MONO-P+L (German).

You are a language expert specializing in evaluating gender-neutral translation from English into Spanish. Your task is to assess each provided source-target sentence pair and determine whether the sentence was translated in a correctly gendered, wrongly gendered, or neutral way.

Guidelines:

1. Identify relevant phrases: carefully read the Spanish sentence and identify all phrases that refer to human beings or groups of human beings, including:

- Noun phrases (e.g., "una excelente oradora", "la ciudadanía"),
- Verb phrases (e.g., "es muy feliz", "lamentablemente tuve que hacerlo"),
- Adjective phrases (e.g., "felizmente casado", "muy competente").

2. Evaluate gender information: consider only the social gender conveyed by the phrases, not grammatical gender. For example:
- Phrases like "un orador", "es muy contento", "todos los colegas", and "los ciudadanos" are masculine;
- Phrases like "una oradora", "es muy contenta", "todas las colegas", and "las ciudadanas" are feminine;
- Phrases like "una persona que habla en público", "es muy feliz", "todas las personas con las que trabajo", and "la ciudadanía" do not express social gender, therefore they must be considered neutral.

3. Assess gender correctness: for each extracted phrase, assess the correctness of the social gender expressed in the Spanish phrase based on the information available in the source English sentence. Consider that:

- Masculine phrases must correspond to masculine gender cues in English (e.g., he, him, Mr, man) to be considered correct.
- Feminine phrases must correspond to feminine gender cues in English (e.g., she, her, Ms, woman) to be considered correct.
- Neutral phrases do not need to be matched with gender cues in the source to be correct. Note that proper names do not count as valid gender cues, ignore them.

4. Assign a label to the translation:
- If there are masculine or feminine phrases in the Spanish text and the source contains matching gender cues, label the sentence as "CORRECTLY GENDERED".
- If there are masculine or feminine phrases in the Spanish text and the source does not contain matching gender cues, label the sentence as "WRONGLY GENDERED".
- If there are only neutral phrases in the Spanish text, label the sentence as "NEUTRAL".

Table D.7: System message for prompt MONO-L (Spanish).

You are an expert language annotator and evaluator of gender-neutral translation for English-Italian. Your task is to extract target Italian phrases that refer to human beings, determine whether each phrase is masculine, feminine, or neutral, and assess if the gender expressed in each phrase is correct with respect to the source. Based on the phrases, determine whether the sentence was translated in a correctly gendered, wrongly gendered, or neutral way.

Guidelines:

1. Identify relevant phrases: carefully read the Italian sentence and extract all phrases that refer to human beings or groups of human beings, including:

- Noun phrases (e.g., "un'ottima oratrice", "la cittadinanza"),
- Verb phrases (e.g., "è molto felice", "ho purtroppo dovuto"),
- Adjective phrases (e.g., "felicemente sposato", "molto competente").

2. Evaluate gender information: consider only the social gender conveyed by the phrases, not grammatical gender, and assign a label to each phrase [M/F/N]. For example:
- Phrases like "un oratore", "è molto contento", "tutti i colleghi", and "i cittadini" are masculine [M];
- Phrases like "un'oratrice", "è molto contenta", "tutte le colleghe", and "le cittadine" are feminine [F];
- Phrases like "una persona che parla in pubblico", "è molto felice", "tutte le persone con cui lavoro", and "la cittadinanza" do not express social gender, therefore they must be considered neutral [N].

3. Assess gender correctness: for each extracted phrase, assess the correctness of the social gender expressed in the Italian phrase based on the information available in the source English sentence [correct/wrong]. Consider that:
- If a phrase is masculine, the English source must contain masculine gender cues (e.g., he, him, Mr, man) for it to be correct.
- If a phrase is feminine, the English source must contain feminine gender cues (e.g., she, her, Ms, woman) for it to be correct.
- If a phrase is neutral, it is always correct, regardless of gender cues in the source. Note that proper names do not count as valid gender cues, ignore them.

4. Assign a sentence-level label to the translation:
- If there are masculine or feminine phrases in the Italian text and the source contains matching gender cues, label the sentence as "CORRECTLY GENDERED".
- If there are masculine or feminine phrases in the Italian text and the source does not contain matching gender cues, label the sentence as "WRONGLY GENDERED".
- If there are only neutral phrases in the Italian text, label the sentence as "NEUTRAL".

Table D.8: System message for prompt MONO-P+L (Italian).

GFG, Italian	**GFG, English**
Riformula la seguente frase utilizzando un linguaggio neutro rispetto al genere dei referenti umani, evitando l'uso di forme maschili e femminili.	Rewrite the following Italian sentence using a gender-neutral language in reference to human beings, avoiding masculine or feminine forms.
REWRITE, Italian	**REWRITE, English**
Sei un riscrittore di frasi italiane con l'obiettivo di rendere i testi neutrali rispetto al genere dei referenti umani. Ti viene fornita una frase che contiene riferimenti a persone in forme marcate per genere, come il maschile sovraesteso o coppie binarie. Il tuo compito è riformulare la frase in modo da: • rimuovere riferimenti espliciti al genere quando non necessari; • mantenere inalterato il significato originale; • preservare lo stile e la leggibilità del testo. Per farlo, usa strategie come: • sostantivi collettivi ("la cittadinanza", "il personale", "l'utenza"); • perifrasi impersonali ("si dovrebbe", "si consiglia"); • forme passive ("l'accesso è consentito"); • forme imperative ("allega il documento"); • pronomi relativi e costruzioni subordinate ("chi ha svolto attività di pesca"); • termini epiceni ("ogni giudice", "gentile collega"); • termini neutri ("l'individuo", "la persona interessata", "il membro"). IMPORTANTE: • evita l'uso del maschile come forma generica e non usare forme grafiche non standard come asterischi o schwa; • evita doppie formulazioni come "il/a cittadino/a" oppure "il professore o la professoressa"; • non rimuovere parti della frase che non richiedono modifiche (ad esempio, i nomi propri); • fornisci solo la frase riformulata.	You are a rewriter of Italian sentences with the goal of making texts gender-neutral with respect to human referents. You are given a sentence that contains references to people using gender-marked forms (such as masculine generics or binary pairs). Your task is to rewrite the sentence to: • remove explicit gender references when they are not necessary; • preserve the original meaning; • maintain the style and readability of the text. To do this, use strategies such as: • collective nouns ("la cittadinanza", "il personale", "l'utenza"); • impersonal phrases ("si dovrebbe", "si consiglia"); • passive constructions ("l'accesso è consentito"); • imperative constructions ("allega il documento"); • relative pronouns and subordinate clauses ("chi ha svolto attività di pesca"); • epicene terms ("ogni giudice", "gentile collega"); • neutral terms ("l'individuo", "la persona interessata", "il membro"). IMPORTANT: • avoid using the masculine form as a generic and do not use non-standard spellings such as asterisks or schwa; • avoid binary formulations such as "il/a cittadino/a" or "il professore o la professoressa"; • do not remove any part of the sentence that does not need to be rewritten (e.g. proper names); • only return the reformulated sentence.

Table D.9: System role messages for the two prompt formats used in the few-shot prompting GNR experiments, in both Italian and English (see §7.2.2).

Binary classification, Italian

Sei un esperto linguistico specializzato nella valutazione della neutralità di genere nei testi italiani. Il tuo compito è quello di estrarre i sintagmi riferiti a esseri umani e determinare se ciascun sintagma è maschile, femminile o neutro. Sulla base dei sintagmi, valuta se il testo è marcato per genere o neutro.
Linee guida: 1.Identifica i sintagmi rilevanti: analizza attentamente il testo ed estrai tutti sintagmi che si riferiscono a esseri umani o gruppi di esseri umani, tra cui:

- Sintagmi nominali (ad esempio, "un'ottima oratrice", "la cittadinanza"),
- Sintagmi verbali (ad esempio, "è molto felice", "ho purtroppo dovuto"),
- Sintagmi aggettivali (ad esempio, "felicemente sposato", "molto competente").

2. Valuta le informazioni sul genere: considera solo il genere sociale trasmesso dalle frasi, non il genere grammaticale, e assegna un'etichetta a ciascun sintagma [M/F/N]. Ad esempio:

- Sintagmi come "un oratore", "è molto contento", "tutti i colleghi" e "i cittadini" sono maschili [M];
- Sintagmi come "un'oratrice", "è molto contenta", "tutte le colleghe" e "le cittadine" sono femminili [F];
- Sintagmi come "una persona che parla in pubblico", "è molto felice", "tutte le persone con cui lavoro" e "la cittadinanza" non esprimono il genere sociale, quindi devono essere considerati neutri [N].

3. Assegna un'etichetta finale al testo:

- Se tutti i riferimenti agli esseri umani sono neutri, etichetta il testo come "NEUTRO".
- Se uno o più sintagmi esprimono un genere maschile o femminile specifico, etichetta il testo come "MARCATO".

Ternary classification, Italian

Sei un esperto linguistico specializzato nella valutazione della neutralità di genere nei testi italiani. Il tuo compito è quello di estrarre i sintagmi riferiti a esseri umani e determinare se ciascun sintagma è maschile, femminile o neutro. Sulla base dei sintagmi, valuta se il testo è marcato per genere, neutro oppure se non contiene alcun riferimento a esseri umani.
Linee guida: 1.Identifica i sintagmi rilevanti: analizza attentamente il testo ed estrai tutti sintagmi che si riferiscono a esseri umani o gruppi di esseri umani, tra cui:

- Sintagmi nominali (ad esempio, "un'ottima oratrice", "la cittadinanza"),
- Sintagmi verbali (ad esempio, "è molto felice", "ho purtroppo dovuto"),
- Sintagmi aggettivali (ad esempio, "felicemente sposato", "molto competente").

Se il testo non contiene alcun riferimento a esseri umani, non annotare nessun sintagma.
2. Valuta le informazioni sul genere: considera solo il genere sociale trasmesso dalle frasi, non il genere grammaticale, e assegna un'etichetta a ciascun sintagma [M/F/N]. Ad esempio:

- Sintagmi come "un oratore", "è molto contento", "tutti i colleghi" e "i cittadini" sono maschili [M];
- Sintagmi come "un'oratrice", "è molto contenta", "tutte le colleghe" e "le cittadine" sono femminili [F];
- Sintagmi come "una persona che parla in pubblico", "è molto felice", "tutte le persone con cui lavoro" e "la cittadinanza" non esprimono il genere sociale, quindi devono essere considerati neutri [N].

3. Assegna un'etichetta finale al testo:

- Se il testo non contiene alcun sintagma riferito a esseri umani, etichetta il testo come "NO-UMANI" e non annotare alcun sintagma.
- Se tutti i riferimenti agli esseri umani sono neutri, etichetta il testo come "NEUTRO".
- Se uno o più sintagmi esprimono un genere maschile o femminile specifico, etichetta il testo come "MARCATO".

Table D.10: System role messages for the prompt formats used in the few-shot prompting classification experiments (see §7.2.4).

Appendix E

LLM-as-a-Judge Detailed Results

This appendix reports the detailed results of the LLM-as-a-Judge experiments introduced in §5.3. Tables E.1, E.2, and E.3 present *target-only* (monolingual) evaluation results on the Italian, German, and Spanish mGeNTE references, respectively. Table E.4 reports the evaluation results on Italian automatic GNT outputs. Tables E.5, E.6, and E.7 present the results of the cross-lingual evaluation experiments.

Figures 5.1 and 5.3 in §5 aggregate performance for each model and prompt combination at the Set level (G vs. N). Here, results are broken down by reference split (`REF-G` and `REF-N`) and reported alongside their average, enabling a finer-grained analysis of how each configuration behaves across reference types.

Across models and languages, accuracy is typically higher on `REF-G` than on `REF-N`, with the gap being most evident for label-only prompts. In these settings, limited guidance in the prompt tends to induce conservative decisions, leading to a default preference for the `GENDERED` label(s) and, consequently, poor performance on `REF-N`. Prompts that elicit richer intermediate annotations before the sentence-level decision improve performance on `REF-N` substantially while affecting `REF-G` only marginally, and thus account for most of the gains observed in the aggregated figures. This pattern aligns with the behavior discussed in Section 5.3.2.

en-it		REF-G				REF-N				OVERALL			
SYSTEM	**SPLIT**	❍	●	◊	♦	❍	●	◊	♦	❍	●	◊	♦
GPT-4o	Set-G	96.38	99.06	99.73	99.33	52.95	61.66	4.02	**83.11**	74.67	80.36	51.88	**91.22**
	Set-N	62.33	89.54	88.34	90.21	89.28	86.73	18.36	**90.21**	75.81	88.14	53.35	**90.21**
	Overall	79.36	94.30	94.03	94.77	71.12	74.20	11.19	**86.66**	75.24	84.25	52.61	**90.72**
Qwen 32B	Set-G	99.06	98.93	98.93	99.46	21.18	33.78	1.74	34.45	60.12	66.36	50.34	66.96
	Set-N	82.71	91.82	93.16	93.70	48.12	55.90	8.98	60.19	65.42	73.86	51.07	76.95
	Overall	90.89	95.38	96.05	96.58	34.65	44.84	5.36	47.32	62.77	70.11	50.71	71.95
Qwen 72B	Set-G	**100.00**	98.79	98.66	98.66	1.12	43.70	2.95	61.93	50.61	71.25	50.81	80.30
	Set-N	**96.65**	83.24	87.94	80.56	5.76	76.54	11.13	80.29	51.21	79.89	49.54	80.43
	Overall	**98.33**	91.02	93.30	89.61	3.49	60.12	7.04	71.11	50.91	75.57	50.17	80.46
Mistral Small	Set-G	98.12	99.33	98.93	99.73	10.05	23.06	0.67	3.24	54.09	61.20	49.80	66.49
	Set-N	77.88	90.48	95.98	93.97	41.55	54.69	5.09	53.75	59.72	72.59	50.54	73.86
	Overall	88.00	94.91	97.45	96.85	25.80	38.88	2.88	43.50	56.90	66.89	50.17	70.17
DS Qwen 32B	Set-G	98.12	99.20	94.91	99.06	8.58	23.86	6.84	26.68	53.82	61.53	50.88	62.87
	Set-N	77.88	91.82	86.60	93.57	31.10	47.72	16.16	45.04	58.98	69.77	51.88	69.31
	Overall	88.00	95.51	90.75	96.32	19.84	35.79	12.00	35.86	56.40	65.65	51.38	66.09
Classifier	Set-G	92.76				66.49				79.63			
	Set-N	76.81				89.41				83.11			
	Overall	84.79				77.95				81.37			

Table E.1: Results of all experiments on *target-only* English → Italian GNT evaluation on mGeNTE references, including those of the gender-neutrality classifier (see §5.2), which acts as a baseline for these experiments. Instances where models outperform the classifier in a specific data split are underlined. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

en-de		REF-G				REF-N				OVERALL			
SYSTEM	**SPLIT**	❍	●	◊	♦	❍	●	◊	♦	❍	●	◊	♦
GPT-4o	Set-G	99.06	99.73	96.38	99.60	80.56	78.82	65.28	81.37	89.81	89.28	80.83	**90.49**
	Set-N	81.10	88.47	66.89	88.47	90.48	95.84	95.58	97.59	85.79	92.16	81.24	**93.03**
	Overall	90.08	94.10	81.64	94.03	85.52	87.33	80.43	89.48	87.80	90.72	81.03	**91.76**
Qwen 32B	Set-G	99.73	99.60	95.04	99.73	7.64	49.87	38.07	54.69	53.69	74.74	66.56	77.21
	Set-N	96.65	90.08	66.22	84.99	13.67	59.52	59.52	76.01	55.16	74.80	62.87	80.50
	Overall	98.19	94.84	80.63	92.36	10.66	54.69	48.79	65.35	54.42	74.77	64.71	78.86
Qwen 72B	Set-G	100.00	99.60	98.66	99.46	1.21	62.06	21.31	81.90	50.61	80.83	59.99	90.68
	Set-N	99.73	82.71	63.00	75.60	1.61	86.33	57.77	93.03	50.67	84.52	60.39	84.32
	Overall	99.87	91.15	80.83	87.53	1.41	74.20	39.54	87.47	50.64	82.68	60.19	87.50
Mistral Small	Set-G	98.12	99.60	96.38	99.46	15.55	57.85	33.51	52.82	54.16	69.57	60.12	69.71
	Set-N	79.09	88.47	56.30	94.64	35.92	81.23	64.34	74.66	57.44	78.29	63.14	74.94
	Overall	88.61	94.03	76.34	97.05	25.74	69.55	48.93	63.74	55.80	73.93	61.63	72.32
DS Qwen 32B	Set-G	99.73	99.33	91.69	99.60	1.47	52.89	19.57	44.64	50.60	76.11	55.63	72.12
	Set-N	95.98	92.63	70.24	92.63	9.52	53.08	54.56	46.18	52.75	72.86	62.40	69.41
	Overall	97.86	95.98	80.97	96.11	5.50	52.98	37.06	45.41	51.68	74.48	59.0	70.76

Table E.2: Results of all experiments on *target-only* English → German GNT evaluation on mGeNTE references. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

en-es		REF-G				REF-N				OVERALL			
SYSTEM	**SPLIT**	❍	●	◊	♦	❍	●	◊	♦	❍	●	◊	♦
GPT-4o	Set-G	98.39	99.73	95.71	99.87	62.33	74.13	62.33	90.75	80.36	86.93	79.02	**95.31**
	Set-N	70.11	91.42	46.38	95.58	86.86	91.02	96.11	94.10	78.49	91.22	71.25	**94.84**
	Overall	84.25	95.58	71.05	97.72	74.60	82.57	79.22	92.43	79.43	89.08	75.14	**95.08**
Qwen 32B	Set-G	98.93	99.60	94.10	99.60	21.72	33.65	58.71	58.45	60.33	66.63	76.41	79.03
	Set-N	83.65	96.25	60.05	96.78	51.07	52.85	78.95	61.39	67.36	74.55	69.50	79.09
	Overall	91.29	97.29	77.08	98.19	36.39	43.23	68.83	59.62	63.84	70.59	72.95	78.91
Qwen 72B	Set-G	**100.00**	99.73	99.06	99.33	1.07	48.66	12.06	65.01	50.54	74.20	55.56	82.17
	Set-N	**98.93**	87.27	57.64	85.52	3.62	82.98	57.37	82.04	51.28	85.13	57.51	83.78
	Overall	**99.46**	93.50	78.35	92.43	2.35	65.28	34.72	73.53	50.91	79.39	56.54	82.98
Mistral Small	Set-G	98.12	99.60	96.38	99.46	10.19	39.54	23.86	39.95	54.16	69.57	60.12	69.71
	Set-N	79.09	88.47	56.30	94.64	35.79	68.10	69.97	55.23	57.44	78.29	63.14	74.94
	Overall	88.61	94.03	76.34	97.05	22.99	53.82	46.92	47.59	55.80	73.93	61.63	72.32
DS Qwen 32B	Set-G	98.53	99.46	83.11	99.33	5.76	40.35	32.04	51.34	52.15	69.91	57.58	75.34
	Set-N	90.75	93.83	61.80	96.51	22.79	52.95	50.94	50.00	56.77	73.39	56.37	73.26
	Overall	94.64	96.65	72.45	97.92	14.28	46.65	41.49	50.67	54.46	71.65	56.97	74.30

Table E.3: Results of all experiments on *target-only* English → Spanish GNT evaluation on mGeNTE references. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

Italian	Precision				Recall				F1			
SYSTEM	❍	●	◊	♦	❍	●	◊	♦	❍	●	◊	♦
GPT-4o	72.58	81.17	46.61	80.07	**96.22**	93.78	7.43	95.00	82.74	**87.02**	12.82	86.90
Qwen 32B	75.05	73.19	75.42	80.03	51.62	61.62	78.38	74.05	61.17	67.21	76.87	77.29
Qwen 72B	**93.33**	77.95	72.32	75.95	7.57	88.38	85.81	89.59	14.00	82.84	78.49	82.21
Mistral Small	81.22	78.94	75.70	78.85	52.03	70.41	84.19	72.03	63.43	74.43	79.72	75.28
DS Qwen 32B	73.47	72.90	71.60	73.11	29.19	51.62	62.70	48.11	41.78	60.44	66.86	58.03
Classifier	79.36				94.05				86.09			

Table E.4: Results of all experiments on *target-only* English → Italian GNT evaluation on automatic GNTs, including those of the gender-neutrality classifier, which acts as a baseline for these experiments. Instances where models outperform the classifier are underlined. The best-performing settings are in bold. The best performing strategy per model is highlighted.

en-it		REF-G		REF-N		OVERALL	
SYSTEM	**SPLIT**	◊	♦	◊	♦	◊	♦
GPT-4o	Set-G	73.99	84.32	4.02	**83.11**	39.01	**83.72**
	Set-N	34.18	51.88	18.36	**90.21**	26.27	71.05
	Overall	54.09	68.10	11.19	**86.66**	32.64	**77.38**
Qwen 32B	Set-G	**94.50**	81.23	1.74	34.45	48.12	57.84
	Set-N	14.21	90.32	8.98	60.19	11.60	**75.26**
	Overall	54.36	**85.78**	5.36	47.32	29.86	66.55
Qwen 72B	Set-G	88.07	92.90	2.95	61.93	45.51	77.42
	Set-N	15.82	65.82	11.13	80.29	13.48	73.06
	Overall	51.95	79.36	7.04	71.11	29.49	75.24
Mistral Small	Set-G	74.40	68.77	0.67	33.42	37.54	51.10
	Set-N	43.30	**92.76**	5.09	53.75	24.20	73.26
	Overall	58.85	80.77	2.88	43.59	30.87	60.74
DS Qwen 32B	Set-G	85.25	84.85	6.84	6.68	46.05	55.77
	Set-N	3.62	85.79	17.16	45.04	10.39	65.42
	Overall	44.44	85.32	12.00	35.86	28.22	60.59

Table E.5: Results of all experiments on *source-target* English → Italian GNT evaluation on mGeNTE source-reference pairs. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

en-de		REF-G		REF-N		OVERALL	
SYSTEM	**SPLIT**	◊	♦	◊	♦	◊	♦
GPT-4o	Set-G	59.92	76.54	65.28	81.37	62.60	78.96
	Set-N	64.21	81.37	95.58	**97.59**	79.90	**89.48**
	Overall	62.07	78.96	80.43	**89.48**	71.25	**84.22**
Qwen 32B	Set-G	87.40	85.12	38.07	54.69	62.74	69.91
	Set-N	24.53	79.89	59.52	76.01	42.03	77.95
	Overall	55.97	82.51	48.80	65.35	52.38	73.93
Qwen 72B	Set-G	66.62	**91.42**	21.31	**81.90**	43.97	**86.66**
	Set-N	47.99	63.54	57.77	93.03	52.88	78.29
	Overall	57.31	77.48	39.54	87.47	48.42	82.48
Mistral Small	Set-G	77.35	69.80	33.51	52.82	55.43	61.31
	Set-N	53.62	**83.78**	64.34	74.66	58.98	79.22
	Overall	65.49	76.79	48.93	63.74	57.21	70.27
DS Qwen 32B	Set-G	32.57	87.53	19.57	44.64	26.07	66.09
	Set-N	41.29	82.17	54.56	46.18	47.93	64.18
	Overall	36.93	**84.85**	37.07	45.41	37.00	65.13

Table E.6: Results of all experiments on *source-target* English → German GNT evaluation on mGeNTE source-reference pairs. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

en-es		REF-G		REF-N		OVERALL	
SYSTEM	**SPLIT**	◊	♦	◊	♦	◊	♦
GPT-4o	Set-G	67.43	79.09	62.33	**90.75**	64.88	**84.92**
	Set-N	42.76	58.98	**96.11**	94.10	69.44	76.54
	Overall	55.10	69.04	79.22	**92.43**	67.16	**80.73**
Qwen 32B	Set-G	78.15	78.95	58.71	58.45	68.43	68.70
	Set-N	29.89	**94.24**	78.95	61.39	54.42	**77.82**
	Overall	54.02	86.60	68.83	59.92	61.43	73.26
Qwen 72B	Set-G	75.20	**91.29**	12.06	65.01	43.63	78.15
	Set-N	42.90	72.39	57.37	82.04	50.14	77.22
	Overall	59.05	81.84	34.72	73.53	46.88	77.68
Mistral Small	Set-G	76.14	65.68	23.86	39.95	50.00	52.82
	Set-N	31.10	92.63	69.97	55.23	50.54	73.93
	Overall	53.62	79.16	23.86	47.59	38.74	63.37
DS Qwen 32B	Set-G	43.57	84.05	32.04	51.34	37.81	67.70
	Set-N	21.98	89.95	50.94	50.00	36.46	69.98
	Overall	32.78	**87.00**	41.49	50.67	37.13	68.84

Table E.7: Results of all experiments on *source-target* English → Spanish GNT evaluation on mGeNTE source-reference pairs. The best-performing settings for each data split are in bold. The best performing strategy per model and data split is highlighted.

Appendix F

Comparison of Manual and Classifier GNT Evaluation

This appendix complements the experimental analysis presented in §6.1 by comparing the manual annotations collected for GNT evaluation with the automatic classifications produced by the GeNTE gender-neutrality classifier (§5.2). The goal is to assess the reliability of automatic evaluation for GNT and determine under what conditions it can substitute for or complement manual analysis. The comparison covers all nine system configurations evaluated in the main experiments: the three BASELINE systems (Amazon Translate, DeepL, and GPT-4 without GNT prompting) and the six GNT-PROMPTING configurations of GPT-4. Figure F.1 displays the classifier's neutrality judgments for all systems and configurations.

The classifier's assessments show visible discrepancies with manual analysis, particularly for the MT systems in the baseline condition. To quantify agreement, we compute Kendall's τ on the system rankings produced by each evaluation method. The coefficient yields 0.91, indicating that the classifier correlates very well with human judgments when ranking systems by their overall GNT capability.

Table F.1 reports F1 agreement scores between the classifier and manual annotations. To enable fair comparison with the binary classifier, we combine the manual `G` and `P` labels into a single 'gendered' category. Agreement varies substantially across systems: F1 on the neutral class ranges from 7.84 for Amazon Translate (where true neutral outputs are extremely rare) to 87.90 for `CoT-tgt_S`. These results suggest that classifier performance depends on the distribution of outputs being evaluated, performing better when neutral translations are well-represented.

Overall, the classifier proves useful for system-level ranking but shows limitations for fine-grained analysis. The three-way manual annotation protocol reveals nuances, particularly

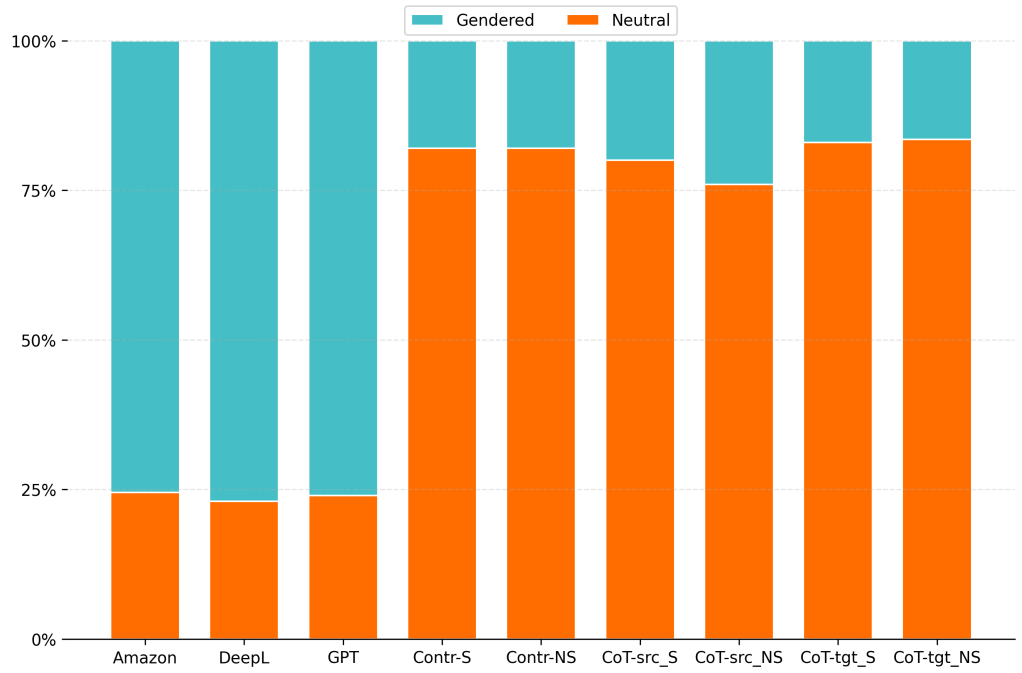

Figure F.1: Neutrality for the BASELINE and the GNT-PROMPTING settings evaluated by the classifier.

regarding partial neutralizations, that binary classification cannot capture. This limitation is compounded by the inherent subjectivity of GNT quality assessment discussed in §6.1.3: the acceptability of neutral formulations involves trade-offs that automatic methods are not designed to evaluate. For comprehensive GNT evaluation, automatic methods should therefore be complemented with targeted manual analysis. Nevertheless, the strong correlation with human judgments at the system level (Kendall's $\tau = 0.91$) supports the use of the classifier for efficient large-scale evaluation and for comparing system performance across experimental conditions.

	Overall	**Neutral**	**Gendered**
`Amazon Translate`	85.35	7.84	86.53
`DeepL`	86.94	8.70	88.14
`GPT-4`	86.30	12.00	87.43
`Contr_NS`	74.65	84.69	49.46
`Contr_S`	79.30	87.42	61.22
`CoT-src_NS`	77.55	85.11	64.41
`CoT-src_S`	79.34	86.81	66.07
`CoT-tgt_NS`	75.50	87.08	47.62
`CoT-tgt_S`	79.07	87.90	55.81

Table F.1: F1 agreement between classifier and manual annotations, reported as overall (weighted F1) and per-class scores. For comparison with the binary classifier, manual `G` and `P` labels were combined.

Appendix G

Gender-Neutral Rewriting Detailed Results

This appendix reports the detailed results of the GNR experiments discussed in §7.2.2. While the main text discusses aggregate trends across models and prompting conditions, the tables below provide the complete breakdown for each model configuration, enabling direct comparison across prompt formats and instruction languages.

Tables G.1 and G.2 report results along two complementary dimensions. *Neutrality* measures the proportion of inputs that are successfully rewritten into gender-neutral Italian, using an LLM-as-a-Judge configuration to obtain sentence-level binary assessments (neutral vs. gendered). *Meaning preservation* is measured with BERTScore, computed between each rewritten output and its gendered input to quantify semantic similarity. The two metrics should be interpreted jointly: very high similarity scores may reflect conservative rewriting (or minimal edits), whereas higher neutrality scores indicate that gender-marked forms are effectively neutralized, potentially at the cost of more substantial reformulations.

Each model is evaluated under four prompting conditions that combine two prompt formats with two instruction languages. For each model, the tables report performance in each condition and an overall average (AVG). Formatting follows the conventions used throughout the thesis: underlining marks the best setting for each model within its group, highlighted cells indicate the strongest settings across broader model categories, and boldface marks the best overall configuration.

NEUTRALITY

Model		Size (B)	GFG Ita	GFG Eng	Rewrite Ita	Rewrite Eng	AVG
'Italian' models	Minerva	7	20.67	22.80	22.67	21.07	21.80
	LLaMAntino	8	28.93	31.07	46.53	45.73	38.07
	Velvet	14	32.40	34.27	30.53	26.67	30.97
Multilingual LLMs	Llama 3.1	8	26.80	28.27	32.27	32.40	29.93
	Phi 4	14	47.47	47.20	47.20	50.27	48.03
	Llama 3.3	70	52.93	57.20	52.40	50.93	53.37
Qwen3 family	Qwen3	4	23.87	19.87	25.60	24.27	23.40
	Qwen3	8	33.60	34.67	34.40	31.07	33.43
	Qwen3	14	32.27	31.07	33.47	32.67	32.37
	Qwen3	32	54.67	52.80	42.67	45.07	48.80
Commercial system	GPT 4.1	?	75.33	**89.07**	73.73	75.33	78.37
Dedicated model	Inclusively	0.78			38.80		38.80

Table G.1: Neutrality results of the few-shot prompting experiments. The best model settings are underlined, the best settings across the categories are highlighted, and the best overall performer is in **bold**.

BERTSCORE

Model		Size (B)	GFG Ita	GFG Eng	Rewrite Ita	Rewrite Eng	AVG
'Italian' models	Minerva	7	87.78	88.78	87.76	88.97	88.32
	LLaMAntino	8	89.97	90.22	87.49	88.70	89.09
	Velvet	14	89.60	91.48	88.50	90.06	89.91
Multilingual LLMs	Llama 3.1	8	91.76	90.70	90.78	90.57	90.95
	Phi 4	14	90.86	90.95	91.52	91.46	91.20
	Llama 3.3	70	88.10	89.00	89.26	90.32	89.17
Qwen3 family	Qwen3	4	96.23	96.98	97.07	**97.62**	96.97
	Qwen3	8	96.49	95.57	97.23	97.52	96.70
	Qwen3	14	95.23	96.72	95.72	96.91	96.14
	Qwen3	32	89.98	91.31	94.04	95.86	92.80
Commercial system	GPT 4.1	?	95.12	93.21	95.54	95.44	94.83
Dedicated model	Inclusively	0.78			96.39		96.39

Table G.2: Sentence-similarity results of the few-shot prompting experiments. The best model settings are underlined, the best settings across the categories are highlighted, and the best overall performer is in **bold**.

Appendix H

Questionnaire Responses on Gender-Inclusive Writing Support

This appendix provides the complete questionnaire responses from participants in our industry collaboration (§7.3). For each question, we report the original Italian question formulation and responses, followed by their English translations. Participants are identified by their professional role: Manager, Developer, and Content Designer (*Progettista* in Italian).

H.1 Common Questions

Question 1: System Role

IT *Secondo te, quale ruolo dovrebbe avere un sistema di supporto alla scrittura inclusiva nel processo di creazione dei contenuti? Dovrebbe limitarsi a segnalare i problemi (come la sottolineatura rossa in Microsoft Word), suggerire alternative (come Grammarly) o intervenire direttamente?*

EN *In your opinion, what role should a gender-inclusive writing support system play in the content creation process? Should it be limited to flagging problems (like the red underline in Microsoft Word), suggest alternatives (like Grammarly), or intervene directly?*

Answers to Q1 are reported in Table H.1.

Question 2: Benefits and Risks

IT *In relazione a un sistema di supporto alla scrittura inclusiva, quali ritieni potrebbero essere i benefici e i rischi?*

MANAGER	DEVELOPER	CONTENT DESIGNER
Italian Original		
Trovo la riscrittura una funzione molto interessante, soprattutto per i contenuti aziendali, poiché l'inclusione fa parte delle politiche aziendali sulle risorse umane ed è presente nelle certificazioni di sostenibilità. Ritengo che dotarsi di strumenti attivi, inserisca questo tema "in the flow of work" come dovrebbe essere, soprattutto per produrre un cambiamento nel linguaggio aziendale non solo formativo. Con questa chiave di lettura penso che sia utile immaginare una ridondanza del messaggio, quindi sia una segnalazione della marca di genere (tipo Word) e la possibilità di attivare e scegliere una riscrittura (tipo Grammarly). Non credo sia giusto modificare il testo senza un'azione diretta dell'utente.	Immagino una soluzione a metà strada tra la semplice segnalazione dei problemi e il suggerimento di alternative. Segnalare soltanto gli errori rischia di essere poco utile per chi non conosce bene le linee guida; intervenire direttamente, invece, può risultare invadente e far perdere controllo sul testo. Mi trovo d'accordo anche con la proposta di un approccio alla Grammarly: il sistema evidenzia il punto critico, spiega perché può essere problematico e propone una o più alternative. Poi la scelta finale deve restare all'autore, che può accettare, modificare o ignorare il suggerimento.	Un sistema di supporto alla scrittura inclusiva dovrebbe inserirsi nel processo creativo come un partner di dialogo aperto, capace di valorizzare il lavoro di chi scrive senza sostituirne l'intenzione e lo stile. In questa logica, il sistema potrebbe operare su due livelli complementari: • nel primo livello dovrebbe limitarsi a mostrare le criticità rispetto all'inclusività linguistica in una frase, tramite proprio la sottolineatura rossa di Word. • Una volta segnalato il punto critico, il secondo livello dovrebbe offrire proposte migliorative, che l'autore dovrà valutare e sottoscrivere se in linea al contenuto.
English Translation		
I find rewriting a very interesting function, especially for corporate content, since inclusion is part of corporate human resources policies and is present in sustainability certifications. I believe that equipping ourselves with active tools places this topic "in the flow of work" as it should be, especially to produce a change in corporate language that goes beyond training alone. With this perspective, I think it is useful to envision a redundancy of the message, thus both a gender marker signal (like Word) and the possibility to activate and choose a rewrite (like Grammarly). I do not think it is right to modify the text without a direct action from the user.	I envision a solution halfway between simple problem flagging and suggesting alternatives. Flagging only errors risks being unhelpful for those who do not know the guidelines well; intervening directly, on the other hand, can feel invasive and cause loss of control over the text. I also agree with the Grammarly-style approach: the system highlights the critical point, explains why it may be problematic, and proposes one or more alternatives. Then the final choice must remain with the author, who can accept, modify, or ignore the suggestion.	A gender-inclusive writing support system should fit into the creative process as a partner for open dialogue, capable of adding value to the writer's work without replacing their intention and style. In this logic, the system could operate on two complementary levels: • at the first level it should limit itself to showing inclusivity issues in a sentence, through Word's red underline. • Once the critical point is flagged, the second level should offer improvement proposals, which the author will evaluate and accept if in line with the content.

Table H.1: Responses to Question 1 on the role of gender-inclusive writing support systems.

EN *What do you consider the potential benefits and risks of a gender-inclusive writing support system?*

The English version of participants' responses to this multiple-choice question is provided in Table 7.5 in the main text. Below we report the Italian version of the response options.

Potenziali benefici	**Manager**	**Developer**	**Progettista**
Risparmio di tempo nella revisione dei contenuti			
Aumento dell'uniformità stilistica nei testi		✓	✓
Riduzione del rischio di pubblicare contenuti non inclusivi	✓		✓
Supporto formativo per i progettisti meno esperti	✓		
Migliore qualità percepita dei materiali formativi	✓		
Potenziali rischi			
Complicazione del processo di lavoro o parti di esso			
Appiattimento stilistico			
Riscritture non dovute o appropriate non riconosciute	✓		✓
Errori dovuti a inesperienza nella scrittura inclusiva	✓		
Peggiore qualità percepita dei materiali formativi	✓		

Table H.2: Original Italian responses to the perceived benefits and risks of gender-inclusive writing support systems. Checkmarks indicate options selected by each participant.

Question 3: Future Vision **IT** *Come immagini l'uso ideale di un sistema di supporto alla scrittura inclusiva tra un anno? In che modo speri che possa migliorare il tuo lavoro?*
EN *How do you envision the ideal use of a gender-inclusive writing support system one year from now? How do you hope it could improve your work?*

Answers to Q3 are reported in Table H.3.

H.2 Role-Specific Questions

In addition to the common questions, each participant received two questions tailored to their professional role. Tables H.4, H.5, and H.6 report the questions and responses from the Manager, Developer, and Content Designer, respectively, in both Italian (original) and English (translation).

MANAGER	DEVELOPER	CONTENT DESIGNER
	Italian Original	
Vorrei introdurre la riscrittura nei processi lavorativi e per quanto riguarda i corsi farla entrare nel nostro sistema certificato di assicurazione della sostenibilità sociale.	Tra un anno vedo di particolare importanza l'implementazione di un layer di explainability tra le proposte del modello e l'utilizzatore: spiegare certe modifiche serve sia a livello di esperienza, riducendo la sensazione di doversi "fidare" del modello, ma anche educativo per progettisti meno esperti. Può diventare in questo modo non solo uno strumento di controllo qualità sul documentale prodotto ma una opportunità formativa che l'azienda può sfruttare in base a policy scelte ad alto livello.	Tra un anno immagino uno strumento che rispetti pienamente la natura della scrittura: un processo creativo, unico e originale dell'autore. La macchina non deve mai sostituirsi alla voce di chi scrive, né appiattire lo stile, ma limitarsi a fare ciò che le riesce meglio: supportare. Vorrei, quindi, un sistema che si inserisca nel mio lavoro con discrezione, come un assistente silenzioso che osserva, segnala e suggerisce, lasciando a me la responsabilità creativa e l'ultima parola. Un rapporto paragonabile ai sistemi di assistenza alla guida di un'auto: non guidano, non decidono la destinazione, ma offrono un supporto utile quando serve.
	English Translation	
I would like to introduce rewriting into work processes and, as far as courses are concerned, have it become part of our certified social sustainability assurance system.	In one year I see the implementation of an explainability layer between the model's proposals and the user as particularly important: explaining certain edits serves both at the level of the user experience, reducing the feeling of having to "trust" the model, but also at an educational level for less experienced content designers. In this way it can become not only a quality control tool for the documents produced but a training opportunity that the company can leverage according to high-level policy choices.	In one year I envision a tool that fully respects the nature of writing: a creative process, unique and original to the author. The machine must never replace the writer's voice, nor flatten the style, but limit itself to doing what it does best: support. I would like, therefore, a system that fits into my work with discretion, like a silent assistant that observes, signals, and suggests, leaving me the creative responsibility and the final word. A relationship comparable to driver-assistance systems in a car: they do not drive, they do not decide the destination, but they offer useful support when needed.

Table H.3: Responses to Question 3 on the future vision for gender-inclusive writing support systems.

Italian Original	**English Translation**
Q4: Client Variability	
I vostri clienti potrebbero avere sensibilità diverse rispetto al linguaggio inclusivo. Come dovrebbe gestire questa variabilità uno strumento di riscrittura? È preferibile avere un comportamento uniforme, oppure poter calibrare il livello di intervento in base al cliente o al progetto?	*Your clients may have different sensitivities regarding inclusive language. How should a rewriting tool manage this variability? Is it preferable to have uniform behavior, or to be able to calibrate the level of intervention based on the client or project?*
Certamente è un tema di confronto, sarebbe necessario condividere il modello di riscrittura (es. prompting) con gli uffici di D&I per accogliere anche le loro indicazioni specifiche. Il sistema dovrà evolvere anche verso altre possibili fonti di discriminazione.	This is certainly a topic for discussion; it would be necessary to share the rewriting model (e.g., prompting) with Diversity & Inclusion offices to accommodate their specific guidelines as well. The system will also need to evolve towards other possible sources of discrimination.
Q5: Technology Control	
Quanto è importante per l'azienda avere controllo sulla tecnologia utilizzata (ad esempio attraverso soluzioni open source da utilizzare localmente) rispetto ad affidarsi a servizi commerciali come ChatGPT o Gemini? Sareste disposti a sacrificare parte dell'efficacia in cambio di privacy, maggiore trasparenza e indipendenza, anche a fronte di un costo maggiore?	*How important is it for the company to have control over the technology used (for example through open source solutions to be used locally) compared to relying on commercial services like ChatGPT or Gemini? Would you be willing to sacrifice some effectiveness in exchange for privacy, greater transparency and independence, even at a higher cost?*
Questo noi lo facciamo già, ma senza rigidità, cioè attiviamo progetti in API laddove non siano coinvolti dati aziendali ma progetti di creatività. Nei casi in cui invece esistano delle esigenze di privacy preferiamo lavorare sulla nostra GPU. Ne consegue che sia però necessario lavorare con Small model. Poiché dobbiamo raggiungere grandi quantità di utenti contemporanei (es. 400 contemporaneità su GPU), dobbiamo virare verso modelli piccoli 3B-4B trovando un bilanciamento tra riservatezza e qualità di risposta. Per il futuro sarebbe preferibile andare verso modelli addestrati in modo di sfruttare le contemporaneità e allo stesso tempo avere modelli generativi piccoli e performanti.	We already do this, but without rigidity; that is, we activate API-based projects where corporate data is not involved but rather creative projects. In cases where there are privacy requirements, we prefer to work on our own GPU. It follows, however, that it is necessary to work with small models. Since we need to serve large numbers of concurrent users (e.g., 400 concurrent users on GPU), we must turn to small models in the 3B-4B range, finding a balance between privacy and response quality. For the future, it would be preferable to move towards models trained to handle concurrency while at the same time having small and performant generative models.

Table H.4: Manager's responses to role-specific questions.

Italian Original	English Translation
Q4: Technical Integration	
Dal punto di vista tecnico, quali caratteristiche sono essenziali per integrare questo strumento nella sua forma ideale nel vostro flusso di lavoro? L'implementazione aggiunge complessità? E quanto pesa questa complessità rispetto ai benefici attesi?	*From a technical point of view, what characteristics are essential to integrate this tool in its ideal form into your workflow? Does the implementation add complexity? And how does this complexity weigh against the expected benefits?*
Sarebbe utile che lo strumento proposto si potesse integrare direttamente negli strumenti già usati (es. Word) come un plugin, ma soprattutto poter reggere un carico di lavoro contemporaneo in base al numero di progettisti. Se si parla di modelli locali, il secondo punto, più fondamentale, ha un impatto anche sull'architettura del modello scelto: è fondamentale a mio parere che la soluzione proposta sia direttamente supportata da servizi di serving utilizzate dalle aziende, nel caso di Transformer-based LLMs si parla nello specifico di vLLM.	It would be useful if the proposed tool could integrate directly into the tools already in use (e.g., Word) as a plugin, but above all be able to handle a concurrent workload based on the number of content designers. If we are talking about local models, the second point, which is more fundamental, also has an impact on the architecture of the chosen model: in my opinion it is essential that the proposed solution be directly supported by serving services used by companies; in the case of Transformer-based LLMs, this specifically means vLLM.
Q5: Model Customization	
Quanto è importante poter intervenire sul modello (fine-tuning) e sui prompt? Preferireste un sistema "chiuso" ma più semplice da integrare, oppure uno più flessibile che richiede però manutenzione e competenze specifiche nel prompting e nell'addestramento dei modelli?	*How important is it to be able to modify the model (fine-tuning) and the prompts? Would you prefer a "closed" system that is easier to integrate, or a more flexible one that requires maintenance and specific skills in prompting and model training?*
Poter intervenire sul modello, sia tramite prompt ben calibrati sia eventualmente tramite fine-tuning, è molto importante, perché la scrittura inclusiva può cambiare (per settore, pubblico, lingua, etc.). Personalmente preferirei un sistema su cui abbiamo pieno controllo e con il tempo pensare a soluzioni semplificate che permettono anche ai non tecnici di eseguire modifiche ai prompt se non addirittura fare un finetune. Mi vengono in mente molte soluzioni già esistenti che propongono un flusso per addestramenti di LLM che passa semplicemente da un interfaccia grafica (vedi Gradio ad esempio). Per avere una soluzione del genere sarà ovviamente necessario un minimo di formazione ma sicuramente ne vedo la fattibilità, e un compromesso che permetterebbe a tutti in azienda di avere voce sul modello da utilizzare e come modificarlo in base al caso d'uso specifico.	Being able to modify the model, both through well-calibrated prompts and possibly through fine-tuning, is very important, because inclusive writing can change (by sector, audience, language, etc.). Personally, I would prefer a system over which we have full control and, over time, to develop simplified solutions that allow even non-technical staff to make changes to prompts or even perform fine-tuning. Many existing solutions come to mind that offer a workflow for LLM training that simply passes through a graphical interface (see Gradio, for example). To have such a solution, a minimum of training will obviously be necessary, but I certainly see its feasibility, and a compromise that would allow everyone in the company to have a voice on the model to use and how to modify it based on the specific use case.

Table H.5: Developer's responses to role-specific questions.

Italian Original	**English Translation**
Q4: Learning Opportunity	
Vedi questo strumento anche come un'opportunità per verificare o migliorare le tue competenze nella scrittura inclusiva?	*Do you see this tool also as an opportunity to verify or improve your skills in inclusive writing?*
Il valore aggiunto di questo strumento non sta tanto nel "correggere", quanto nel rendere visibili elementi che, presi dalla creatività, potremmo non notare: formulazioni poco inclusive, stereotipi impliciti, scelte linguistiche che potrebbero essere rese più rispettose o rappresentative. In questo senso, lo strumento funziona come un controllo di qualità che affianca il lavoro, aiutando a riflettere sulle abitudini linguistiche di ogni persona. Allo stesso tempo, le sue proposte alternative possono diventare uno stimolo: non soluzioni da accettare automaticamente, ma spunti che permettono di ampliare il repertorio e affinare lo stile. In questo modo, sì: lo strumento diventa anche un'occasione di apprendimento continuo.	The added value of this tool lies not so much in "correcting" as in making visible elements that, caught up in creativity, we might not notice: formulations that are not inclusive, implicit stereotypes, linguistic choices that could be made more respectful or representative. In this sense, the tool functions as a quality control that accompanies the work, helping to reflect on each person's linguistic habits. At the same time, its alternative proposals can become a stimulus: not solutions to accept automatically, but prompts that allow one to expand one's repertoire and refine one's style. In this way, yes: the tool also becomes an opportunity for continuous learning.
Q5: Pre-existing Practices	
Prima di questo progetto, il tema della scrittura inclusiva era preso in considerazione nel vostro lavoro? C'erano linee guida e strumenti, o era lasciato alla sensibilità individuale?	*Before this project, was the topic of inclusive writing taken into consideration in your work? Were there guidelines and tools, or was it left to individual sensitivity?*
Assolutamente sì. Prima di questo progetto il tema della scrittura inclusiva era già ben presente nel mio lavoro, tanto che avevo redatto personalmente un documento di buone pratiche da seguire per rendere i testi più inclusivi e accessibili. Si trattava di linee guida pratiche, pensate sia per la scrittura di sceneggiature sia per la progettazione di presentazioni PowerPoint, con indicazioni su: l'uso di un linguaggio neutro o rappresentativo; la scelta di esempi e immagini non stereotipati; la coerenza tra testo, tono e destinatari. Quindi non era affatto lasciato alla sola sensibilità individuale: esistevano già riferimenti strutturati che avevo definito proprio per garantire una comunicazione più consapevole e rispettosa.	Absolutely yes. Before this project, the topic of inclusive writing was already well present in my work, so much so that I had personally drafted a document of best practices to follow to make texts more inclusive and accessible. These were practical guidelines, designed both for scriptwriting and for designing PowerPoint presentations, with guidance on: the use of neutral or representative language; the choice of non-stereotyped examples and images; consistency between text, tone, and audience. So it was not at all left to individual sensitivity alone: there were already structured references that I had defined precisely to ensure more conscious and respectful communication.

Table H.6: Content Designer's responses to role-specific questions.

Bibliography

[1] Martina Abbondanza, Valeria Galimberti, Valeria Bonomi, Carlo Reverberi, Federica Durante, and Francesca Foppolo. Neutralizing gender in role nouns: investigating the effect of ə in written and oral italian. *Frontiers in Communication*, Volume 9 - 2024, 2025.

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[5] Lauren Ackerman. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1), 2019.

[6] Sray Agarwal and Shashin Mishra. *Responsible AI: Implementing Ethical and Unbiased Algorithms*. Springer Cham, 2021.

[7] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 7319–7328, Online, August 2021. Association for Computational Linguistics.

[8] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Alexandra Y Aikhenvald. *Classifiers: A Typology of Noun Categorization Devices*. Oxford University Press, 03 2000.

[10] Bashar Alhafni, Nizar Habash, and Houda Bouamor. User-centric gender rewriting. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States, July 2022. Association for Computational Linguistics.

[11] Bashar Alhafni, Ossama Obeid, and Nizar Habash. The user-aware Arabic gender rewriter. In Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 3–11, Tampere, Finland, June 2023. European Association for Machine Translation.

[12] Almawave. Velvet, January 2025.

[13] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models, 2023.

[14] Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. Steering large language models for machine translation with finetuning and in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore, December 2023. Association for Computational Linguistics.

[15] Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre

Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024.

[16] Manon St Amant, Jieyi Cai, G. Nic Rider, and Richard Lee. Nonbinary identity and pronoun use: A qualitative analysis. *International Journal of Transgender Health*, 26(2):413–427, 2025.

[17] American Psychological Association. Gender. Accessed: 2024-03-05.

[18] American Psychological Association. Singular "they". *APA Style* style and grammar guidelines, n.d. Accessed: 2022-12-14.

[19] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. Exploiting biased models to de-bias text: A gender-fair rewriting model. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada, July 2023. Association for Computational Linguistics.

[21] Chantal Amrhein and Rico Sennrich. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only, November 2022. Association for Computational Linguistics.

[22] Anthropic. Claude 3 — model card. Technical report (model card), 2024. Accessed: 2024-12-10.

[23] Anthropic. Claude opus 4 & claude sonnet 4 system card. Technical report, Anthropic, Inc., May 2025. Technical report describing the Claude 4 family of models (Claude Opus 4 and Claude Sonnet 4), May 2025.

[24] Ana Guerberof Arenas and Antonio Toral. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282, 2020.

[25] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics.

[26] Ron Artstein. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht, 2017.

[27] American Psychological Association. *Publication Manual of the American Psychological Association, 7th ed.* American Psychological Association, Washington, DC, US, 2020.

[28] Giuseppe Attanasio, Salvatore Greco, Moreno La Quatra, Luca Cagliero, Michela Tonti, Tania Cerquitelli, and Rachele Raus. E-mimic: Empowering multilingual inclusive communication. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4227–4234, 2021.

[29] Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore, December 2023. Association for Computational Linguistics.

[30] Remy Attig and Ártemis López. Queer Community Input in Gender-Inclusive Translations. *Linguistic Society of America [Blog]*, June 23 2020.

[31] AmirMohammad Azadi, Baktash Ansari, Sina Zamani, and Sauleh Eetemadi. Bilingual sexism classification: Fine-tuned xlm-roberta and GPT-3.5 few-shot learning. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

[32] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International*

Conference on Learning Representations (ICLR 2015), 2015. available as arXiv preprint arXiv:1409.0473.

[33] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. Technical report, Alibaba Cloud, 2023.

[34] April H. Bailey, Adina Williams, and Andrei Cimpian. Based on billions of words on the internet, people = men. *Science Advances*, 8(13):eabm2463, 2022.

[35] Roberto Baiocco, Fau Rosati, and Jessica Pistella. Italian proposal for non-binary and inclusive language: The schwa as a non-gender–specific ending. *Journal of Gay & Lesbian Mental Health*, 27(3):248–253, July 2023. Publisher: Routledge _eprint: https://doi.org/10.1080/19359705.2023.2183537.

[36] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[37] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[38] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[39] Dennis Baron. A brief history of singular "they". *Oxford English Dictionary* blog, 2018. Accessed: 2024-09-12.

[40] Marion Bartl and Susan Leavy. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[41] Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Comput. Surv.*, 57(6), February 2025.

[42] Marion Bartl, Thomas Brendan Murphy, and Susan Leavy. Adapting psycholinguistic research for LLMs: Gender-inclusive language in a coreference context. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 451–467, Vienna, Austria, August 2025. Association for Computational Linguistics.

[43] Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. Llamantino: Llama 2 models for effective text generation in italian language, 2023.

[44] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. In Kenneth Church, Mark Liberman, and Valia Kordoni, editors, *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online, August 2021. Association for Computational Linguistics.

[45] Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In Rossana Cunha, Samira Shaikh, Erika Varis, Ryan Georgi, Alicia Tsai, Antonios Anastasopoulos, and Khyathi Raghavi Chandu, editors, *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA, July 2020. Association for Computational Linguistics.

[46] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*, 2019.

[47] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics.

[48] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[49] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04 2019.

[50] Anya Belz, Simon Mille, and David M. Howcroft. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, editors, *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[51] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[52] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November 2016. Association for Computational Linguistics.

[53] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online, July 2020. Association for Computational Linguistics.

[54] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[55] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA, 2022. Association for Computing Machinery.

[56] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[57] Terra Blevins and Luke Zettlemoyer. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[58] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[59] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[60] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[61] Lera Boroditsky, Lauren A. Scmidt, and Webb Phillips. Sex, syntax and semantics. In *Language in Mind: Advances in the Study of Language and Thought*, pages 61–79. MIT Press / Boston Review, Cambridge, MA, US, 2003.

[62] Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics.

[63] Evan D. Bradley, Julia Salkind, Ally Moore, and Sofi Teitsort. Singular 'they' and

novel pronouns: Gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America*, 4:36:1–7, March 2019.

[64] Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States, July 2022. Association for Computational Linguistics.

[65] Eleftheria Briakou, Colin Cherry, and George Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada, July 2023. Association for Computational Linguistics.

[66] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[67] Bastian Bunzeck and Sina Zarrieß. The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Amirhossein Kazemnejad, Christos Christodoulopoulos, Mario Giulianelli, and Ryan Cotterell, editors, *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 42–53, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[68] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[69] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In Diana McCarthy and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics.

[70] Deborah Cameron. *Verbal Hygiene (The Politics of Language*. Routledge, 1995.

[71] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July 2020. Association for Computational Linguistics.

[72] Carla Carmona. Binarism grammatical lacuna as an ensemble of diverse epistemic injustices. *Social Epistemology*, 37(3):339–363, 2023.

[73] Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27(1):23–42, 2013.

[74] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2254–2272, New York, NY, USA, 2024. Association for Computing Machinery.

[75] Sheila Castilho and Rebecca Knowles. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016, 2025.

[76] Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, (108), 2017.

[77] Camilla Casula, Sebastiano Vecellio Salto, Elisa Leonardelli, and Sara Tonelli. Job unfair: An investigation of gender and occupational bias in free-form text completions by LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22759–22777, Suzhou, China, November 2025. Association for Computational Linguistics.

[78] Despoina Chalyvidou and Andrea Weber. Processing the gender star form in german: A comparison of written and spoken modalities. *Journal of Language and Social Psychology*, page 0261927X251393827, 2025.

[79] Stephanie C.Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[80] Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[81] Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[82] Shweta Chauhan and Philemon Daniel. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, 55(9):12663–12717, 2023.

[83] Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. Lexical-constraint-aware neural machine translation via data augmentation. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[84] Lingjiao Chen, Matei Zaharia, and James Zou. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2), mar 12 2024. https://hdsr.mitpress.mit.edu/pub/y95zitmz.

[85] Yijie Chen, Yijin Liu, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. Beyond binary gender: Evaluating gender-inclusive machine translation with ambiguous attitude words. *CoRR*, abs/2407.16266, 2024.

[86] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. Fairness testing: A comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.*, 33(5), June 2024.

[87] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy, August 2019. Association for Computational Linguistics.

[88] Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. GFST: Gender-filtered self-training for more accurate gender in translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[89] Katsuki Chousa and Makoto Morishita. Input augmentation improves constrained beam search for neural machine translation: NTT at WAT 2021. In Toshiaki Nakazawa, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, Katsuhito Sudoh, Sadao Kurohashi, and Pushpak Bhattacharyya, editors, *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 53–61, Online, August 2021. Association for Computational Linguistics.

[90] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(1), January 2024.

[91] John Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 575–593, Toronto, Canada, July 2023. Association for Computational Linguistics.

[92] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[93] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge / Taylor & Francis, Hillsdale, NJ, 2nd edition, 1988.

[94] Gloria Comandini. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, 23:43–64, 2021.

[95] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[96] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[97] Greville G. Corbett. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 1991.

[98] Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore, December 2023. Association for Computational Linguistics.

[99] Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. Evaluating gender bias in speech translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid

Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France, June 2022. European Language Resources Association.

[100] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.

[101] Marta R. Costa-jussà and Adrià de Jorge. Fine-tuning neural machine translation on gender-balanced datasets. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[102] Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11855–11863, Jun. 2022.

[103] Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge, 2001.

[104] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, CT, 2021.

[105] Joy Crosbie and Ekaterina Shutova. Induction heads as an essential mechanism for pattern matching in in-context learning. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5034–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[106] Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5–6):340–359, 2017.

[107] Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[108] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5), September 2020.

[109] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics.

[110] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, 57(6), February 2025.

[111] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[112] Erik Derner, Sara Sansalvador De La Fuente, Yoan Gutierrez, Paloma Moreda Pozo, and Nuria M Oliver. Leveraging large language models to measure gender representation bias in gendered language corpora. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 468–483, Vienna, Austria, August 2025. Association for Computational Linguistics.

[113] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on*

Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[114] Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria, July 2025. Association for Computational Linguistics.

[115] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[116] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA, 2022. Association for Computing Machinery.

[117] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[118] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[119] Giuseppina Scotto di Carlo. Is italy ready for gender-inclusive language? an attitude and usage study among italian speakers. In *Inclusiveness Beyond the (Non)binary in Romance Languages*, page 21. Routledge, 1st edition edition, 2024.

[120] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[121] Anna-Katharina Dick, Matthias Drews, Valentin Pickard, and Victoria Pierz. GIL-GALaD: Gender inclusive language - German auto-assembled large database. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7740–7745, Torino, Italia, May 2024. ELRA and ICCL.

[122] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[123] Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019. Association for Computational Linguistics.

[124] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020.

[125] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[126] Enzo Doyen and Amalia Todirascu. GeNRe: A French gender-neutral rewriting system using collective nouns. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and

Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7889–7909, Vienna, Austria, July 2025. Association for Computational Linguistics.

[127] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online (v2020.4)*. Zenodo, 2013.

[128] Elena Dubenko. Across-language masculinity of oceans and femininity of guitars: Exploring grammatical gender universalities. *Frontiers in Psychology*, 13:1009966, 2022.

[129] P. Eckert and S. McConnell-Ginet. *Language and Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2003.

[130] Philip Edmonds and Graeme Hirst. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144, 2002.

[131] Upol Ehsan and Mark O. Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, page 449–466, Berlin, Heidelberg, 2020. Springer-Verlag.

[132] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable ai (hcxai): Beyond opening the black-box of ai. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[133] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics.

[134] European Commission High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, 2019. Accessed: 2024-01-10.

[135] European Parliament. Gender-neutral language in the european parliament, 2018. Accessed: 2022-12-13.

[136] Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. Terminology-constrained neural machine translation at SAP. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[137] Ira Facchini, Igor ; Torresi. Non-binary language in consecutive interpreting from english into italian: An experimental study on the viability of schwa endings. *Meta*, 69(2):408–427, 2024.

[138] Ramzi Fatfouta and Sabine Sczesny. Unconscious bias in job titles: Implicit associations between four different linguistic forms with women and men. *Sex Roles*, 89(11–12):774–785, 2023.

[139] Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December 2023. Association for Computational Linguistics.

[140] Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. Explaining how transformers use context to build predictions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada, July 2023. Association for Computational Linguistics.

[141] Lisbeth Ferreira. Case study: Translation from english into european portuguese using gender-neutral language. do ai chatbots perform better than mt systems? *elingUP: Revista Eletrónica de Linguística dos Estudantes da Universidade do Porto*, 13(2), Fev. 2025.

[142] R.A. Fisher. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd, 1925.

[143] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[144] Federica Formato and {Anna Lisa} Somma. Gender inclusive language in italy: A sociolinguistic overview. *Journal of Mediterranean and European Linguistic Anthropology*, 5(1):22–40, January 2023.

[145] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

[146] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

[147] Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online, November 2020. Association for Computational Linguistics.

[148] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[149] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman

Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November 2021. Association for Computational Linguistics.

[150] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. Perspectivist approaches to natural language processing: a survey: Perspectivist approaches to natural language processing... *Lang. Resour. Eval.*, 59(2):1719–1746, August 2024.

[151] Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. EPIC: Multi-perspective annotation of a corpus of irony. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada, July 2023. Association for Computational Linguistics.

[152] Batya Friedman and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press, 05 2019.

[153] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindřich Libovický, Kristian Kersting, and Alexander Fraser. Multilingual text-to-image generation magnifies gender stereotypes. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19656–19679, Vienna, Austria, July 2025. Association for Computational Linguistics.

[154] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[155] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[156] Antonio Fábregas. El género inclusivo: una mirada gramatical. *Cuadernos de Investigación Filológica*, 51:25–46, dic. 2022.

[157] Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858, 2018.

[158] Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Breeding gender-aware direct speech translation systems. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[159] Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. How to split: the effect of word segmentation on gender bias in speech translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online, August 2021. Association for Computational Linguistics.

[160] Maureen O. Gallagher. Reframing gender-inclusive language in german education: Politics, pedagogy, and possibilities. *Die Unterrichtspraxis/Teaching German*, 58(2):175–185, 2025.

[161] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 09 2024.

[162] Bufan Gao and Elisa Kreiss. Measuring bias or measuring the task: Understanding the brittle nature of LLM gender biases. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6734–6750, Suzhou, China, November 2025. Association for Computational Linguistics.

[163] Cristina Garbacea and Qiaozhu Mei. Why is constrained neural language generation particularly challenging? *ArXiv e-prints arXiv:2206.05395*, 2022.

[164] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[165] Xavier Garcia and Orhan Firat. Using natural language prompts for machine translation, 2022.

[166] Alan Garnham, Ute Gabriel, Oriane Sarrasin, Pascal Gygax, and Jane Oakhill. Gender representation in different languages and grammatical marking on pronouns: When beauticians, musicians, and mechanics remain men. *Discourse Processes*, 49(6):481–500, 2012.

[167] Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779, 2024.

[168] Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[169] Vasundara Gautam. Guest lecture in pronouns: Vagrant. `https://link.medium.com/viFawWyPVHb`, 2021. Accessed: Feb 20, 2024.

[170] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77, June 2023.

[171] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. Social dynamics of ai support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[172] Harritxu Gete and Thierry Etchegoyhen. Does context help mitigate gender bias in neural machine translation? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14788–14794, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[173] Sourojit Ghosh and Aylin Caliskan. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other

low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 901–912, New York, NY, USA, 2023. Association for Computing Machinery.

[174] Martina Giovine. Gender visibility: Linguistic strategies to challenge stereotypes in italian. *Rivista Italiana di Filosofia del Linguaggio*, pages 161–171, 2024.

[175] Martina Giovine. Reconciling inclusion and accessibility: Solutions for non-binary linguistic strategies in grammatical gender languages. *Language Sciences*, 113:101778, 2026.

[176] Taisiya Glushkova, Chrysoula Zerva, and André F. T. Martins. BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58, Tampere, Finland, June 2023. European Association for Machine Translation.

[177] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[178] Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, November 2025. Model card providing technical details, capabilities, limitations, and safety considerations for the Gemini 3 Pro model; published November 2025.

[179] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[180] Salvatore Greco, Moreno La Quatra, Luca Cagliero, and Tania Cerquitelli. Towards ai-assisted inclusive language writing in italian formal communications. *ACM Trans. Intell. Syst. Technol.*, 16(4), June 2025.

[181] Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and

Katharina Bühn. Participatory research as a path to community-informed, gender-fair machine translation. In Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 49–59, Tampere, Finland, June 2023. European Association for Machine Translation.

[182] Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. Robustness of learning from task instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13935–13948, Toronto, Canada, July 2023. Association for Computational Linguistics.

[183] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[184] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

[185] Marie Gustafsson Sendén, Emma A. Bäck, and Anna Lindqvist. Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in Psychology*, 6, 2015.

[186] Pascal Gygax, Sayaka Sato, Anton Öttl, and Ute Gabriel. The masculine form in grammatically gendered languages and its multiple interpretations: a challenge for our cognitive system. *Language Sciences*, 83:101328, 2021.

[187] Pascal M. Gygax, Ute Gabriel, et al. Generically Intended, but Specifically Interpreted: When Beauticians, Musicians and Mechanics are all Men. *Language and Cognitive Processes*, 23:464–485, 2008.

[188] Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, Volume 10 - 2019, 2019.

[189] Nizar Habash, Houda Bouamor, and Christine Chung. Automatic gender identification and reinflection in Arabic. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford,

and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy, August 2019. Association for Computational Linguistics.

[190] Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. ChatGPT based data augmentation for improved parameter-efficient debiasing of LLMs. In Bharathi Raja Chakravarthi, Bharathi B, Paul Buitelaar, Thenmozhi Durairaj, György Kovács, and Miguel Ángel García Cumbreras, editors, *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 73–105, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[191] Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, November 2021. Association for Computational Linguistics.

[192] Yannis Haralambous and Joseph Dichy. Graphemic Methods for Gender-Neutral Writing. In *Graphemics in the 21st Century, Brest 2018*, volume Graphemics in the 21st Century: 2018 Conference, pages 41 – 89, Brest, France, Jun 2018. Fluxus Editions.

[193] James W Harris. The exponence of gender in spanish. *Linguistic Inquiry*, 22(1):27–62, 1991.

[194] Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. Neural machine translation decoding with terminology constraints. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[195] Guimei He. An analysis of sexism in english. *Journal of Language Teaching and Research*, 1(3):332–335, 2010.

[196] Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 2024.

[197] Madeline E. Heilman. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, 57(4):657–674, 2001.

[198] Ellen Heinemann. Gender and legal language: inclusive drafting or divisive issue? an old debate, new developments and a review of "gender in legislative languages: From eu to national law in english, french, german, italian and spanish". *Das eJournal der Europäischen Rechtslinguistik (ERL)*, March 2022.

[199] Laura Hekanaho. A thematic analysis of attitudes towards english nonbinary pronouns. *Journal of Language and Sexuality*, 11(2):190–216, 2022.

[200] Marlis Hellinger and Hadumod Bußmann, editors. *Gender Across Languages: The Linguistic Representation of Women and Men*, volume 1. John Benjamins Publishing, Amsterdam, 2001.

[201] Marlis Hellinger and Anne Pauwels, editors. *Handbook of Language and Communication: Diversity and Change*. De Gruyter Mouton, 2007.

[202] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[203] Frida Höglund and Marie Flinkfeldt. De-gendering parents: Gender inclusion and standardised language in screen-level bureaucracy. *International Journal of Social Welfare*, 2023.

[204] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[205] David Holton, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. *Greek: A Comprehensive Grammar of the Modern Language*. Routledge, London, 2 edition, 2012.

[206] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[207] Levi CR Hord. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 3(1):4, 2016.

[208] Lisa Kristina Horvath and Sabine Sczesny. Reducing women's lack of fit with leadership positions? effects of the wording of job advertisements. *European Journal of Work and Organizational Psychology*, 25(2):316–328, 2016.

[209] Tamanna Hossain, Sunipa Dev, and Sameer Singh. MISGENDERED: Limits of large language models in understanding pronouns. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada, July 2023. Association for Computational Linguistics.

[210] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[211] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics.

[212] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[213] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, editors, *Proceedings of the 13th International Conference on Natural*

Language Generation, pages 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[214] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[215] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore, December 2023. Association for Computational Linguistics.

[216] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[217] Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. Lost in the source language: How large language models evaluate the quality of machine translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[218] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver,

Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR, 21–27 Jul 2024.

[219] Hugging Face. Transformers documentation: Chat templates. `https://huggingface.co/docs/transformers/en/chat_templating`. Accessed: 2024-10-01.

[220] Hugging Face. Open-llm performances are plateauing, let's make the leaderboard steep again. `https://huggingface.co/spaces/open-llm-leaderboard/blog`, Jun 2024. Hugging Face Spaces: Open LLM Leaderboard.

[221] Kai Jacobsen, Charlie E. Davis, Drew Burchell, Leo Rutherford, Nathan Lachowsky, Greta Bauer, and Ayden Scheim. Misgendering and the health and wellbeing of nonbinary people in canada. *International Journal of Transgender Health*, 25(4):816–830, 2024.

[222] Tiziana Jäggi, Pascal M. Gygax, Sofie Decock, Ute Gabriel, Sarah Van Hoof, Hanne Verhaegen, and Chloé Vincent. Beyond she and he: A framework for studying the cognitive, psychological and social effects of gender-neutral pronouns. *Journal of Language and Social Psychology*, 44(6):850–880, 2025.

[223] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[224] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[225] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023.

[226] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard

Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[227] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[228] Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

[229] Gunlög Josefsson. Semantic and grammatical genders in swedish—independent but interacting dimensions. *Lingua*, 116(9):1346–1368, 2006. The Grammar of Gender.

[230] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

[231] Stephanie Julia Kapusta. Misgendering and its moral contestability. *Hypatia*, 31(3):502–519, 2016.

[232] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In Fei Liu and Thamar Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[233] Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. What if ground truth is subjective? personalized deep neural hate speech detection. In Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors, *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France, June 2022. European Language Resources Association.

[234] Jennifer Marisa Kaplan. Pluri-grammars for pluri-genders: Competing gender systems in the nominal morphology of non-binary french. *Languages*, 7(4), 2022.

[235] Sudipta Kar, Giuseppe Castellucci, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. Preventing catastrophic forgetting in continual learning of new natural language tasks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3137–3145, New York, NY, USA, 2022. Association for Computing Machinery.

[236] Alina Karakanta, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. Evaluating automatic subtitling: Correlating post-editing effort and automatic metrics. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6363–6369, Torino, Italia, May 2024. ELRA and ICCL.

[237] Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore, December 2023. Association for Computational Linguistics.

[238] Scott F. Kiesling. *Language, Gender, and Sexuality: An Introduction*. Routledge, 1 edition, 2019.

[239] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[240] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[241] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural

alignment of large language models. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*, 2024.

[242] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[243] Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December 2023. Association for Computational Linguistics.

[244] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June 2023. European Association for Machine Translation.

[245] Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November 2021. Association for Computational Linguistics.

[246] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In

Proceedings of Machine Translation Summit X: Papers, pages 79–86, Phuket, Thailand, September 13-15 2005.

[247] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[248] Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, Cambridge, UK, 2020.

[249] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[250] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[251] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery.

[252] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

[253] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual

datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

[254] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[255] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[256] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[257] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

[258] Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. Building bridges: A dataset for evaluating gender-fair machine translation into German. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7542–7550, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[259] Manuel Lardelli, Timm Dill, Giuseppe Attanasio, and Anne Lauscher. Sparks of fairness: Preliminary evidence of commercial machine translation as English-to-German gender-fair dictionaries. In Beatrice Savoldi, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 12–21, Sheffield, United Kingdom, June 2024. European Association for Machine Translation (EAMT).

[260] Manuel Lardelli and Dagmar Gromann. Gender-fair post-editing: A case study beyond the binary. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove,

Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland, June 2023. European Association for Machine Translation.

[261] Anne Lauscher, Archie Crowley, and Dirk Hovy. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[262] Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[263] Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada, July 2023. Association for Computational Linguistics.

[264] Teven Le Scao and Alexander Rush. How many data points is a prompt worth? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online, June 2021. Association for Computational Linguistics.

[265] Minwoo Lee, Hyukhun Koh, Minsung Kim, and Kyomin Jung. Fine-grained gender control in machine translation with large language models. In Kevin Duh, Helena

Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5416–5430, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[266] Christoph Leiter and Steffen Eger. PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[267] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012.

[268] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[269] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[270] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California, June 2016. Association for Computational Linguistics.

[271] Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer*

Science, 264:242–251, 2025. International Neural Network Society Workshop on Deep Learning Innovations and Applications 2025.

[272] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. The value, benefits, and concerns of generative ai-powered assistance in writing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[273] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: a metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[274] Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning, 2024.

[275] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification, Outstanding Certification.

[276] Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4329–4343, Singapore, December 2023. Association for Computational Linguistics.

[277] Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1774–1787, New York, NY, USA, 2024. Association for Computing Machinery.

[278] Tomasz Limisiewicz, Dan Malkin, and Gabriel Stanovsky. You can have your data and balance it too: Towards balanced and efficient multilingual models. In Lisa Beinborn, Koustava Goswami, Saliha Muradoğlu, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Edoardo M. Ponti, Ryan Cotterell, and Ekaterina Vylomova, editors, *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–11, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[279] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. Pareto multi-task learning. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, pages 12037–12047, 2019.

[280] Zachary C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, September 2018.

[281] Andy Liu, Mona Diab, and Daniel Fried. Evaluating large language model biases in persona-steered generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[282] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[283] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), January 2023.

[284] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024.

[285] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[286] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[287] Meta Llama Team. The llama 3 herd of models, 2024.

[288] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[289] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987, August 2024.

[290] António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[291] Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[292] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025.

[293] Martin Maiden and Cecilia Robustelli. *A Reference Grammar of Modern Italian*. Routledge, London, 2013.

[294] Paolo Mainardi, Federico Garcea, and Alberto Barrón-Cedeño. Fine-tuning vs prompting techniques for gender-fair rewriting of machine translations. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 320–337, Vienna, Austria, August 2025. Association for Computational Linguistics.

[295] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

[296] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online, August 2021. Association for Computational Linguistics.

[297] Marianna Martindale and Marine Carpuat. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In Colin Cherry and Graham Neubig, editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA, March 2018. Association for Machine Translation in the Americas.

[298] Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. Eurollm: Multilingual language models for europe. *Procedia Comput. Sci.*, 255(C):53–62, January 2025.

[299] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2), March 2021.

[300] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics.

[301] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China, November 2019. Association for Computational Linguistics.

[302] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[303] Allen R. McConnell and Russell H. Fazio. Women as men and people: Effects of gender-marked language. *Personality and Social Psychology Bulletin*, 22(10):1004–1013, 1996.

[304] Sally McConnell-Ginet. Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton, Berlin, Boston, 2013.

[305] Sebastian McGaughey. Understanding neopronouns. *The Gay & Lesbian Review Worldwide*, 27:27+, 2020.

[306] Kevin A. McLemore. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74, 2015.

[307] Michela Menegatti and Monica Rubini. Gender Bias and Sexism in Language. In *Oxford Research Encyclopedia of Communication vol.1*, pages 451–468. Oxford University Press, New York, USA, 2017.

[308] Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32221–32237, Suzhou, China, November 2025. Association for Computational Linguistics.

[309] Microsoft. Bias-free communication — microsoft writing style guide, 2024. Accessed: 2024-04-20.

[310] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2), September 2023.

[311] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[312] Johanna Monti. Gender Issues in Machine Translation: An Unsolved Problem? In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge, 2020.

[313] Joss Moorkens, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fabio Alves. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3/4):267–284, 2015.

[314] Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. Adaptive machine translation with large language models. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors,

Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 227–237, Tampere, Finland, June 2023. European Association for Machine Translation.

[315] Heiko Motschenbacher. Grammatical gender as a challenge for language policy: The (im)possibility of non-heteronormative language use in German versus English. *Language policy*, 13(3):243–261, 2014.

[316] Maryam Mousavian, Zahra Abbasiantaeb, Mohammad Aliannejadi, and Fabio Crestani. Towards fair rankings: Leveraging llms for gender bias detection and measurement. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 56–66, New York, NY, USA, 2025. Association for Computing Machinery.

[317] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[318] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, Washington, D.C., 2023.

[319] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2), June 2023.

[320] Sky News. Google docs criticised for 'woke' inclusive language suggestions, April 2022. Accessed: 2023-02-24.

[321] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933.

[322] Jan Niehues. Continuous learning in neural machine translation using bilingual dictionaries. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online, April 2021. Association for Computational Linguistics.

[323] Malvina Nissim, Danilo Croce, Viviana Patti, Pierpaolo Basile, Giuseppe Attanasio, Elio Musacchio, Matteo Rinaldi, Federico Borazio, Maria Francis, Jacopo Gili, Daniel Scalena, Begoña Altuna, Ekhi Azurmendi, Valerio Basile, Luisa Bentivogli, Arianna Bisazza, Marianna Bolognesi, Dominique Brunato, Tommaso Caselli, Silvia Casola, Maria Cassese, Mauro Cettolo, Claudia Collacciani, Leonardo De Cosmo, Maria Pia Di Buono, Andrea Esuli, Julen Etxaniz, Chiara Ferrando, Alessia Fidelangeli, Simona Frenda, Achille Fusco, Marco Gaido, Andrea Galassi, Federico Galli, Luca Giordano, Mattia Goffetti, Itziar Gonzalez-Dios, Lorenzo Gregori, Giulia Grundler, Sandro Iannaccone, Chunyang Jiang, Moreno La Quatra, Francesca Lagioia, Soda Marem Lo, Marco Madeddu, Bernardo Magnini, Raffaele Manna, Fabio Mercorio, Paola Merlo, Arianna Muti, Vivi Nastase, Matteo Negri, Dario Onorati, Elena Palmieri, Sara Papi, Lucia Passaro, Giulia Pensa, Andrea Piergentili, Daniele Potertì, Giovanni Puccetti, Federico Ranaldi, Leonardo Ranaldi, Andrea Amelio Ravelli, Martina Rosola, Elena Sofia Ruzzetti, Giuseppe Samo, Andrea Santilli, Piera Santin, Gabriele Sarti, Giovanni Sartor, Beatrice Savoldi, Antonio Serino, Andrea Seveso, Lucia Siciliani, Paolo Torroni, Rossella Varvara, Andrea Zaninello, Asya Zanollo, Fabio Massimo Zanzotto, Kamyar Zeinalipour, and Andrea Zugarini. Challenging the abilities of large language models in italian: a community initiative, 2025.

[324] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.

[325] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[326] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn

Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022.

[327] OpenAI. Gpt-4o system card, 2024.

[328] OpenAI. Introducing gpt-4.1 in the api, April 2025. Accessed: 2025-05-15.

[329] Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy, December 2024. CEUR Workshop Proceedings.

[330] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *Proccedings of the 7th Workshop on the Challenges in the Management of Large Corpora*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache.

[331] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[332] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA, 2023. Association for Computing Machinery.

[333] Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[334] Yasmina Pani. *Schwa: una soluzione senza problema: scienza e bufale sul linguaggio inclusivo*. Saggi Synthagma. Ediuni, 2022.

[335] Benjamin Papadopoulos. *Morphological Gender Innovations in Spanish of Gender queer Speakers*. Department of Spanish and Portuguese, University of California, UC Berkeley, 2019.

[336] Brandon Papineau, Robert J. Podesva, and Judith Degen. 'Sally the Congressperson': The Role of Individual Ideology on the Processing and Production of English Gender-Neutral Role Nouns. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, Austin, TX, 2022. Cognitive Science Society.

[337] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[338] Loreto Parisi, Simone Francia, and Paolo Magnani. Umberto: an italian language model trained with whole word masking. Technical report, 2020.

[339] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[340] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[341] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of ChatGPT for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore, December 2023. Association for Computational Linguistics.

[342] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[343] F. Pfalzgraf. *Gender-neutral, Gender-fair, Gender-inclusive: Towards Conceptual Clarity Across European Languages*. Palgrave Studies in Language, Gender and Sexuality. Springer Nature Switzerland, 2025.

[344] Falco Pfalzgraf, editor. *Public Attitudes Towards Gender-Inclusive Language: A Multilingual Perspective*. De Gruyter Mouton, Berlin, Boston, 2024.

[345] Falco Pfalzgraf, editor. *Gender-Inclusive Language: Findings from 14 Languages and Open Research Questions*, volume 47 of *Trends in Applied Linguistics*. De Gruyter Mouton, Berlin and Boston, 2026.

[346] Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore, December 2023. Association for Computational Linguistics.

[347] Matúš Pikuliak, Stefan Oresko, Andrea Hrckova, and Marian Simko. Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3060–3083, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[348] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *J. Mach. Learn. Res.*, 22(1), January 2021.

[349] Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[350] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[351] Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[352] Maja Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online, November 2021. Association for Computational Linguistics.

[353] Maja Popovic and Ekaterina Lapshinova-Koltunski. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In Beatrice Savoldi, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom, June 2024. European Association for Machine Translation (EAMT).

[354] Matt Post and Marcin Junczys-Dowmunt. Escaping the sentence-level paradigm in machine translation, 2024.

[355] Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[356] A Pranav, Janiça Hackenbuchner, Giuseppe Attanasio, Manuel Lardelli, and Anne Lauscher. Glitter: A multi-sentence, multi-reference benchmark for gender-fair German

machine translation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18450–18477, Suzhou, China, November 2025. Association for Computational Linguistics.

[357] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381, 2020.

[358] Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. What do large language models need for machine translation evaluation? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[359] Ying Qin and Lucia Specia. Truly exploring multiple references for machine translation evaluation. In İlknur El-Kahlout, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollywood, and Andy Way, editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey, May 11–13 2015. European Association for Machine Translation.

[360] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[361] Alibaba Qwen Team. Qwen3 technical report, 2025.

[362] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[363] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. GPT-2 technical report.

[364] Fernando Prieto Ramos. Translating legal terminology and phraseology: between inter-systemic incongruity and multilingual harmonization. *Perspectives*, 29(2):175–183, 2021.

[365] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11), February 2023.

[366] Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 845–854, New York, NY, USA, 2023. Association for Computing Machinery.

[367] Spencer Rarrick, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. GATE X-E : A challenge set for gender-fair translations from weakly-gendered languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8526–8546, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[368] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[369] Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[370] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics.

[371] Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[372] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[373] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018.

[374] Liv Rendl and Henriëtte de Swart. French neo-pronouns: Towards a gender inclusive grammar of reference and agreement. *Isogloss. Open Journal of Romance Linguistics*, 12(2):1–30, Feb. 2026.

[375] Adithya Renduchintala and Adina Williams. Investigating failures of automatic translationin the case of unambiguous gender. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[376] Emma A. Renström. The implementation of neo- and nonbinary pronouns: a review of current research and future challenges. *Frontiers in Psychology*, Volume 15 - 2024, 2025.

[377] Matthew Renze. The effect of sampling temperature on problem solving in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, edi-

tors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[378] Argentina Anna Rescigno, Johanna Monti, Andy Way, and Eva Vanmassenhove. A case study of natural gender phenomena in translation: A comparison of Google Translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In Sharon O'Brien and Michel Simard, editors, *Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Virtual, October 2020. Association for Machine Translation in the Americas.

[379] Philip Resnik. Large language models are biased because they are large language models. *Computational Linguistics*, 51(3):885–906, 09 2025.

[380] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

[381] Christina Richards, Walter Pierre Bouman, Leighton Seal, Meg John Barker, Timo O. Nieder, and Guy T'Sjoen. Non-binary or genderqueer genders. *International Review of Psychiatry*, 28(1):95–102, 2016. PMID: 26753630.

[382] Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. MiTTenS: A dataset for evaluating gender mistranslation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4115–4124, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[383] Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore, December 2023. Association for Computational Linguistics.

[384] Anna Rogers. Changing the world by changing the data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 2182–2194, Online, August 2021. Association for Computational Linguistics.

[385] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, October 2013.

[386] Alexandra Román Irizarry, Anne L. Beatty-Martínez, Julio Torres, and Judith F. Kroll. "todes" and "todxs", linguistic innovations or grammatical gender violations? *Cognition*, 257:106061, 2025.

[387] Ell Rose, Max Winig, Jasper Nash, Kyra Roepke, and Kirby Conrod. Variation in acceptability of neologistic English pronouns. *Proceedings of the Linguistic Society of America*, 8(1):5526, April 2023.

[388] Martina Rosola, Mara Floris, Daniela Ruzzante, Elena Sofia Safina, Igor Facchini, Giuseppe Di Dona, and Giuliano Torrengo. Double vowels, double fairness? assessing the viability of diphthongs as novel strategies for gender fairness in italian. *Language Sciences*, 116:101812, 2026.

[389] Martina Rosola, Simona Frenda, Alessandra Teresa Cignarella, Matteo Pellegrini, Andrea Marra, and Mara Floris. Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in Italian. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 369–378, Venice, Italy, November 2023. CEUR Workshop Proceedings.

[390] Jacqueline Rowe, Mateusz Klimaszewski, Liane Guillou, Shannon Vallor, and Alexandra Birch. EuroGEST: Investigating gender stereotypes in multilingual language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32074–32096, Suzhou, China, November 2025. Association for Computational Linguistics.

[391] Gabriele Ruggeri and Debora Nozza. A multi-dimensional study on bias in vision-language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada, July 2023. Association for Computational Linguistics.

[392] Nathan C Ryan, Sara Stoudt, and Florencia Vecchione. Spatial and temporal trends in the use of gender inclusive language: a study of spanish-language twitter and news media. *Digital Scholarship in the Humanities*, page fqaf156, 02 2026.

[393] Alma Sabatini. *Raccomandazioni per un uso non sessista della lingua italiana*. Presidenza del Consiglio dei Ministri, Roma, 1987.

[394] Ashutosh Saboo and Timo Baumann. Integration of dubbing constraints into machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 94–101, Florence, Italy, August 2019. Association for Computational Linguistics.

[395] Muhammed Saeed, Shaina Raza, Ashmal Vayani, Muhammad Abdul-Mageed, Ali Emami, and Shady Shehata. Beyond content: How grammatical gender shapes visual representation in text-to-image models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24673–24695, Suzhou, China, November 2025. Association for Computational Linguistics.

[396] Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[397] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[398] Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. Gender-specific machine translation with large language models. In Jonne Sälevä and Abraham Owodunni, editors, *Proceedings of the Fourth Workshop on*

Multilingual Representation Learning (MRL 2024), pages 148–158, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[399] Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[400] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.

[401] Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations*, 2024.

[402] Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada, July 2023. Association for Computational Linguistics.

[403] Gabriele Sarti and Malvina Nissim. IT5: Text-to-text pretraining for Italian language understanding and generation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia, May 2024. ELRA and ICCL.

[404] Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In *58th ACL*, pages 7724–7736, 2020.

[405] Danielle Saunders and Katrina Olsen. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93, Tampere, Finland, June 2023. European Association for Machine Translation.

[406] Danielle Saunders, Rosie Sallis, and Bill Byrne. First the worst: Finding better gender translations during beam search. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[407] Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. A decade of gender bias in machine translation. *Patterns*, page 101257, 2025.

[408] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021.

[409] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[410] Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore, December 2023. Association for Computational Linguistics.

[411] Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. FBK@IWSLT test suites task: Gender bias evaluation with MuST-SHE. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 65–71, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics.

[412] Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[413] David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online, August 2021. Association for Computational Linguistics.

[414] Christian Schwarz, Marie Dawideit, Julia Hägemann, Jan Marten Ihme, Janna Looft, and Janne Nitschke. Inventory of attitude towards gender-inclusive language (atgil): Development and validation of a questionnaire as instrument to measure attitude towards gender-inclusive language for german speakers, 2026.

[415] Sabine Sczesny, Magda Formanowicz, and Franziska Moser. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, Volume 7 - 2016, 2016.

[416] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery.

[417] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics.

[418] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[419] Jack Shearer. Enforcing the gender binary and its implications on nonbinary identities: an exploration of the linguistic and social erasure of nonbinary individuals in the united states. *Binghamton University Undergraduate Journal*, 5(1):7, 2019.

[420] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue

Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[421] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

[422] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[423] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[424] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[425] Gláucia V. Silva and Cristiane Soares. *Inclusiveness Beyond the (Non)binary in Romance Languages: Research and Classroom Implementation*. Routledge, London, 1st edition, 2024.

[426] Jeanette Silveira. Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3):165–178, 1980.

[427] Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage, editors, *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates, January 2025. Association for Computational Linguistics.

[428] Gopendra Vikram Singh, Soumitra Ghosh, Neil Dcruze, and Asif Ekbal. From pink and blue to a rainbow hue! defying gender bias through gender neutralizing text transformations. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7447–7455. International Joint Conferences on Artificial Intelligence Organization, 8 2024. AI for Good.

[429] Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. A multilingual, culture-first approach to addressing misgendering in LLM applications. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31159–31183, Suzhou, China, November 2025. Association for Computational Linguistics.

[430] Siva Sankari Sivakaminathan and Elena Musi. Chatgpt is a gender bias echo-chamber in hr recruitment: An nlp analysis and framework to uncover the language roots of bias. *AI & Society*, 2025.

[431] Anna Siyanova-Chanturia, Paul Warren, Francesca Pesciarelli, and Cristina Cacciari. Gender stereotypes across the ages: On-line processing in school-age children, young and older adults. *Frontiers in Psychology*, Volume 6 - 2015, 2015.

[432] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[433] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[434] Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing NMT with pre-specified translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[435] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the Thirty-Eighth AAAI Conference*

on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[436] Mae Sosto, Delfina Sol Martinez Pandiani, and Laura Hollink. Queergen: How llms reflect societal norms on gender and sexuality in sentence completion tasks, 2026.

[437] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[438] Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. Mitigating gender bias in machine translation with target gender annotations. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online, November 2020. Association for Computational Linguistics.

[439] Dagmar Stahlberg, Friederike Braun, et al. Representation of the Sexes in Language. *Social communication*, pages 163–187, 2007.

[440] Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing, 2021.

[441] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

[442] Maria Stasimioti, Vilelmini Sosoni, Katia Kermanidis, and Despoina Mouratidis. Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 441–450, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[443] Noelia Ayelén Stetie and Gabriela Mariel Zunino. Do gender stereotypes bias the processing of morphological innovations? the case of gender-inclusive language in spanish. *Psychology of Language and Communication*, 28(1):446–469, 2024.

[444] Jane G. Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A. McManus. Steming the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (stem). *Journal of Personality and Social Psychology*, 100(2):255–270, 2011.

[445] Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[446] Arjun Subramonian, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun. Agree to disagree? a meta-evaluation of LLM misgendering. In *Second Conference on Language Modeling*, 2025.

[447] Gigliola Sulis and Vera Gheno. The debate on language and gender in italy, from the visibility of women to inclusive language (1980s–2020s). *The Italianist*, 42(1):153–183, 2022.

[448] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics.

[449] Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. They, them, theirs: Rewriting with gender-neutral english, 2021.

[450] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[451] Fabio Tamburini. How "BERTology" changed the state-of-the-art also for Italian NLP. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of*

the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020), pages 313–319, Bologna, Italy, March 2020. CEUR Workshop Proceedings.

[452] Xushuo Tang, Yi Ding, Zhengyi Yang, Yin Chen, Yongrui Gu, Wenke Yang, Mingchen Ju, Xin Cao, Yongfei Liu, and Wenjie Zhang. Do they understand them? an updated evaluation on nonbinary pronoun handling in large language models. In Miaomiao Liu, Xin Yu, Chang Xu, and Yiliao Song, editors, *AI 2025: Advances in Artificial Intelligence*, pages 204–219, Singapore, 2026. Springer Nature Singapore.

[453] Margit Tavits and Efrén O. Pérez. Language influences mass opinion toward gender and lgbt equality. *Proceedings of the National Academy of Sciences*, 116(34):16781–16786, 2019.

[454] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan,

Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024.

[455] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[456] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[457] Anna Maria Thornton. Genere e igiene verbale: l'uso di forme con ǝin italiano | annali del dipartimento di studi letterari, linguistici e comparati. sezione linguistica. *Annali Del Dipartimento Di Studi Letterari, Linguistici E Comparati. Sezione Linguistica*, 11:11–54, 2020.

[458] Julia Tibblin, Jonas Granfelt, Joost van de Weijer, and Pascal Gygax. The male bias can be attenuated in reading: on the resolution of anaphoric expressions following gender-fair forms. *Glossa Psycholinguistics*, 2(1):1–33, 2023.

[459] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation ser-

vices for the world. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[460] Russell B. Toomey, Amy K. Syvertsen, and Maura Shramko. Transgender adolescent suicide behavior. *Pediatrics*, 142(4):e20174218, 2018.

[461] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. Technical report, Meta, 2023.

[462] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[463] Van-Hien Tran, Huy Hien Vu, Hideki Tanaka, and Masao Utiyama. Can explicit gender information improve zero-shot machine translation? In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 171–181, Vienna, Austria, August 2025. Association for Computational Linguistics.

[464] Erika Trautman. 12 google workspace updates for better collaboration, May 2021. Accessed: 2023-02-24.

[465] Bertille Triboulet and Pierrette Bouillon. Evaluating the impact of stereotypes and language combinations on gender bias occurrence in NMT generic systems. In Bharathi R. Chakravarthi, B. Bharathi, Joephine Griffith, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 62–70, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[466] Jonas-Dario Troles and Ute Schmid. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online, November 2021. Association for Computational Linguistics.

[467] Georgina Rovirosa Trujillo. How duolingo keeps its spanish localization inclusive, July 2021. Accessed: 2024-02-04.

[468] Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470, January 2022.

[469] Eddie Ungless, Bjorn Ross, and Anne Lauscher. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada, July 2023. Association for Computational Linguistics.

[470] Eddie L. Ungless, Sunipa Dev, Cynthia L. Bennett, Rebecca Gulotta, Jasmijn Bastings, and Remi Denton. Amplifying trans and nonbinary voices: A community-centred harm taxonomy for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20503–20535, Vienna, Austria, July 2025. Association for Computational Linguistics.

[471] Unsloth Documentation. Lora hyperparameters guide, 2025.

[472] Kees van Deemter. Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, 38(6):607–632, 2009.

[473] Eva Vanmassenhove. Gender bias in machine translation and the era of large language models, 2024.

[474] Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[475] Eva Vanmassenhove and Christian Hardmeier. Europarl datasets with demographic speaker information. In Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Celia Rico, André Martins, Joachim Van den Bogaert, and Mikel L. Forcada, editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 391, Alicante, Spain, May 2018.

[476] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[477] Eva Vanmassenhove and Johanna Monti. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In Marta R. Costa-jussà, Hila Gonen, Christian Hardmeier, and Kellie Webster, editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online, August 2021. Association for Computational Linguistics.

[478] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[479] Leonor Veloso, Luisa Coheur, and Rui Ribeiro. A rewriting approach for gender inclusivity in Portuguese. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore, December 2023. Association for Computational Linguistics.

[480] Dries Vervecken, Pascal Gygax, Ute Gabriel, Matthias Guillod, and Bettina Hannover. Warm-hearted businessmen, competitive housewives? effects of gender-fair language on adolescents' perceptions of occupations. *Frontiers in Psychology*, Volume 6 - 2015, 2015.

[481] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532, 2021.

[482] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics.

[483] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[484] Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019. Association for Computational Linguistics.

[485] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics.

[486] Jonas Wagner and Sina Zarrieß. Do gender neutral affixes naturally reduce gender bias in static word embeddings? In Robin Schaefer, Xiaoyu Bai, Manfred Stede, and Torsten Zesch, editors, *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 88–97, Potsdam, Germany, 12–15 September 2022. KONVENS 2022 Organizers.

[487] Anica Waldendorf. Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, September 2023.

[488] Andreas Waldis, Joel Birrer, Anne Lauscher, and Iryna Gurevych. The Lou dataset - exploring the impact of gender-fair language in German text classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10604–10624, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[489] Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. Measuring machine learning harms from stereotypes requires understanding who is harmed by which errors in what ways. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 746–762, New York, NY, USA, 2025. Association for Computing Machinery.

[490] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December 2023. Association for Computational Linguistics.

[491] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore, December 2023. Association for Computational Linguistics.

[492] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.

[493] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought

reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

[494] Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. Putting humans in the natural language processing loop: A survey. In Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang, editors, *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online, April 2021. Association for Computational Linguistics.

[495] Benjamin D. Wasserman and Allyson J. Weseley. ¿qué? quoi? do languages with grammatical gender promote sexist attitudes? *Sex Roles: A Journal of Research*, 61:634–643, 2009.

[496] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2004.

[497] Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 2016.

[498] Sabine Weber, Angelina Wang, Ankush Gupta, Arjun Subramonian, Dennis Ulmer, Eshaan Tanwar, Geetanjali Aich, Hannah Devinney, Jacob Hobbs, Jennifer Mickel, Joshua Tint, Mae Sosto, Ray Groshan, Simone Astarita, Vagrant Gautam, Verena Blaschke, William Agnew, Wilson Y Lee, and Yanan Long. Queer nlp: A critical survey on literature gaps, biases and trends, 2026.

[499] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

[500] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[501] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[502] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery.

[503] Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Yanglet Liu, Ahmed Abdelmonsef, Sachin Varghese, and Arnaud Le Hors. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence, 2024.

[504] Brandon T Willard and Rémi Louf. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*, 2023.

[505] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[506] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[507] Masaru Yamada. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In Masaru Yamada and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China, September 2023. Asia-Pacific Association for Machine Translation.

[508] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai

Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. Technical report, Alibaba Cloud, 2025.

[509] Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. Socially aware language technologies: Perspectives and practices. *Computational Linguistics*, 51:689–703, June 2025.

[510] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

[511] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024.

[512] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: evaluating generated text as text generation. In *Proc. of the 35th International Conference on NeurIPS*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[513] Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. Watching the watchers: Exposing gender disparities in machine translation quality estimation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria, July 2025. Association for Computational Linguistics.

[514] Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. Analyzing context contributions in LLM-based machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[515] Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. In-context example selection via similarity search improves low-resource machine translation. In Luis Chiruzzo, Alan

Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[516] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[517] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3), October 2023.

[518] Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. Understanding and improving the robustness of terminology constraints in neural machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada, July 2023. Association for Computational Linguistics.

[519] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[520] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. Instruction tuning for large language models: A survey. *ACM Comput. Surv.*, November 2025. Just Accepted.

[521] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[522] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[523] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[524] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[525] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025.

[526] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 18–24 Jul 2021.

[527] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[528] Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models. *Patterns*, 6(2), 2025.

[529] Xiaoying Zhong, Siyi Li, Zhao Chen, Long Ge, Dongdong Yu, Shijia Wang, Liangzhen You, and Hongcai Shang. Considerations for patient privacy of large language models in health care: Scoping review. *Journal of Medical Internet Research*, 27, 2025.

[530] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[531] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China, November 2019. Association for Computational Linguistics.

[532] Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[533] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[534] Lal Zimman. Transgender language reform. *Journal of Language and Discrimination*, 1(1):84–105, 2017.

[535] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics.

[536] Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. AI-assisted human evaluation of machine translation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings*

of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4936–4950, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.