UNIVERSITY OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING
FONDAZIONE BRUNO KESSLER

Doctorate Program in Industrial Innovation

Empathetic Text Style Transfer Layer for Conversational Agents

A modular architecture and new resource for the Italian language

Simone Manai

Advisor
Dr. Alberto Lavelli
Fondazione Bruno Kessler

Co-Advisor
PhD. Laura Gemme
Lutech Softjam

24 April 2026

Abstract

The integration of Emotional Intelligence (EI) into Artificial Intelligence (AI) systems represents a central challenge in the advancement of Human-Computer Interaction (HCI), particularly in domains such as healthcare where communication requires both factual accuracy and appropriate emotional attenuation. Despite advances in natural language generation, current conversational agents rarely exhibit genuine empathetic behavior, often producing neutral or affectively inadequate responses. This limitation is even more evident for the Italian language, where the development of empathy-aware systems is constrained by the scarcity of high-quality annotated resources and the absence of parallel corpora for controlled stylistic generation. Moreover, incorporating empathetic capabilities directly into existing domain-specific conversational agents entails significant risks for system reliability and imposes high retraining and maintenance costs. To address these constraints, this thesis introduces a modular framework based on an Empathetic Text Style Transfer. This architectural paradigm decouples the generation of empathetic expression from the underlying dialogic and domain-expert components, enabling the enhancement of pre-existing conversational agent infrastructures without modifying their internal logic or compromising factual robustness. A key contribution of this work is the construction and rigorous human evaluation of IDRE (Italian Dialogue for Empathetic Responses), a novel dataset consisting of curated triplets—user query, neutral response, and empathetic reformulation specifically designed to support supervised training for empathetic text style transfer while guaranteeing semantic fidelity. This research bridges a critical gap in Italian NLP resources by introducing IDRE, the first corpus for empathetic text style transfer in this language. Extensive experimental analysis comparing ten distinct model architectures reveals that fine-tuning on the IDRE dataset significantly outperforms few-shot learning strategies, enabling even compact models (approximately 1 billion parameters) to achieve high levels of lexical diversity and emotional resonance. The results confirm that this modular capability is not only robust within the medical domain but generalizes effectively to financial, legal, and social contexts. Ultimately, this work provides a scalable, cost-effective, and computationally efficient framework for deploying human-centric AI, offering a sustainable alternative to resource-intensive commercial models.

Keywords

[Empathetic AI, Large Language Models, IDRE Dataset, Fine-tuning, Human-Computer Interaction]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General framework and motivation

The integration of emotional intelligence into artificial intelligence systems constitutes a pivotal frontier in the evolution of human–computer interaction [19, 77, 63, 40]. In particular, the development of conversational agents capable of expressing empathy is attracting increasing attention due to its potential to support user-centric, trustworthy, and emotionally adaptive interactions.

To properly situate this research within the technological landscape, it is essential to outline the broader domain of Conversational AI [45]. This field encompasses a continuum of systems that differ in architectural complexity, reasoning capabilities, and operational constraints. The literature typically distinguishes the following categories [36, 2, 58]:

- Chatbots: Historically the most basic iteration, these systems operate on predefined rules or keyword-matching logic. They are typically task-oriented and domain-specific, designed to handle linear interactions (e.g., FAQ retrieval) within rigid boundaries.

- Virtual Assistants (VAs): These represent a more sophisticated tier, often characterized by a broader functional scope and the ability to manage multi-turn dialogues. VAs (e.g., Siri, Alexa) act as personal facilitators, leveraging natural language processing to execute tasks across various domains.

- Intelligent Virtual Agents (IVAs): These are enterprise-grade evolution of chatbots that utilize Advanced Natural Language Understanding (NLU) and Machine

Learning. Unlike basic chatbots, IVAs can handle complex transactions and maintain context over extended interactions, often integrating with backend corporate infrastructures.

- Conversational Agents: The most recent paradigm, powered by Large Language Models (LLMs), capable of open-domain reasoning and high linguistic fluidity, though often lacking the deterministic control required in sensitive vertical sectors.

Regardless of the specific architectural paradigm adopted, the integration of affective capabilities requires a rigorous theoretical definition of the phenomenon. Therefore, to ground the design of the proposed solution, it is essential to first analyze the construct of empathy from a scientific perspective.

Empathy is a multidimensional construct that encompasses both the ability to understand and the capacity to resonate with another person's emotional state while maintaining a clear distinction between self and other. In the scientific literature, empathy is typically divided into two main components: affective empathy, involving the sharing of emotional experiences, and cognitive empathy, referring to the accurate understanding of another's emotions [16].

Neuroscientific studies by Decety and Jackson (2004) have demonstrated that empathy engages specific brain circuits, including the medial prefrontal cortex, anterior insula, and anterior cingulate cortex—regions activated both during the direct experience of pain and when observing pain in others [17]. These findings support the embodied simulation hypothesis, which posits that the brain internally simulates the emotional states observed in others, thereby facilitating intersubjective understanding [23].

In developmental psychology, Hoffman (2000) introduced an evolutionary model of empathy, highlighting its early emergence in children and its progression through increasingly sophisticated stages—from global affective responses to cognitively mediated empathic understanding [26]. Carl Rogers, a pioneer of client-centered psychotherapy, defined it as the ability to perceive another's internal frame of reference "as if" one were the other person, while maintaining the crucial "as-if" distinction that preserves self–other boundaries [60]. Within medicine, empathy is essential to effective clinical practice: it enhances the physician–patient relationship, improves adherence to treatment, and reduces patient anxiety. A pivotal study in Academic Medicine found that higher levels of perceived physician empathy correlate with better clinical outcomes, particularly in chronic pain management [27]. Reflecting its importance, the SPIKES protocol—widely adopted in oncology—explicitly includes empathy as a core step in the

delivery of difficult news [6]. Moreover, empathy plays a protective role for healthcare providers: recent longitudinal evidence from the U.S. Veterans Health Administration shows a negative correlation between empathic capacity and burnout rates [47], while a meta-analysis links physician burnout to decreased patient safety and reduced professionalism [49].

Given this extensive body of evidence, empathetic conversational agents hold substantial promise for enhancing user experience in diverse scenarios, including customer support, education, and mental health assistance. Their capacity to offer emotionally attuned responses becomes especially relevant in situations of psychological vulnerability. For example, a neutral answer to an anxious user – "*It's important to stay calm and find a way to relax*" may convey useful information yet fail to validate the person's emotional state. Conversely, an empathetic reformulation such as "*I understand that you're feeling anxious right now, and it is okay to feel this way. Let's take a moment to focus on something that can help you relax*" demonstrates emotional attunement and validation, which are central to the therapeutic potential of empathetic communication. This distinction underscores the importance of integrating affective and cognitive empathy mechanisms into conversational agents, thereby fostering trust, emotional safety, and user engagement.

Despite their potential, the direct development and deployment of fully empathetic AI systems present significant technical, economic, and organizational challenges. Many institutions have already invested substantial resources in creating domain-specific chatbots or coversational agents deeply integrated into their operational infrastructure. Modifying or replacing these systems to incorporate empathetic capabilities would not only incur considerable additional costs but also pose risks of performance degradation and operational instability. This issue is particularly pronounced in vertical domains such as legal advisory, financial consulting, healthcare triage, and technical support, where chatbots are meticulously fine-tuned to deliver accurate and context-sensitive information. Embedding empathy into these models would require extensive retraining using domain-specific empathetic datasets, which are often scarce and can compromise the original task performance of the system. Furthermore, the end-to-end development of a new empathetic conversational agent is resource-intensive, involving multiple stages—data acquisition and annotation, model training and validation, system integration, and regulatory compliance testing. For organizations operating under budgetary constraints or lacking access to high-performance infrastructure, such an undertaking may be economically and logistically infeasible.

1.2 Solution and Innovative Aspects

To address these limitations, the thesis proposes a modular architecture based on an Empathetic Text Style Transfer layer. Implemented through a Large Language Model, the layer processes the output of an existing conversational agent and reformulates it into a semantically equivalent response enriched with empathetic linguistic markers. By decoupling the generation of empathetic expression from the agent's core dialogue mechanisms, the architecture offers a non-intrusive and low-risk strategy for enhancing legacy systems with affective capabilities. This approach improves user engagement and fosters more nuanced, human-centric interactions without altering the intent or informational content of the original response.

As illustrated in Figure 1.1, the proposed architecture provides a cost-effective and and easily integrable solution. Operating independently from the conversational agent's core components, the empathy layer preserves factual accuracy and domain-specific relevance while augmenting the emotional resonance of each output. This design ensures that the factual accuracy and domain-specific relevance of the original response are preserved, while enhancing its emotional resonance.

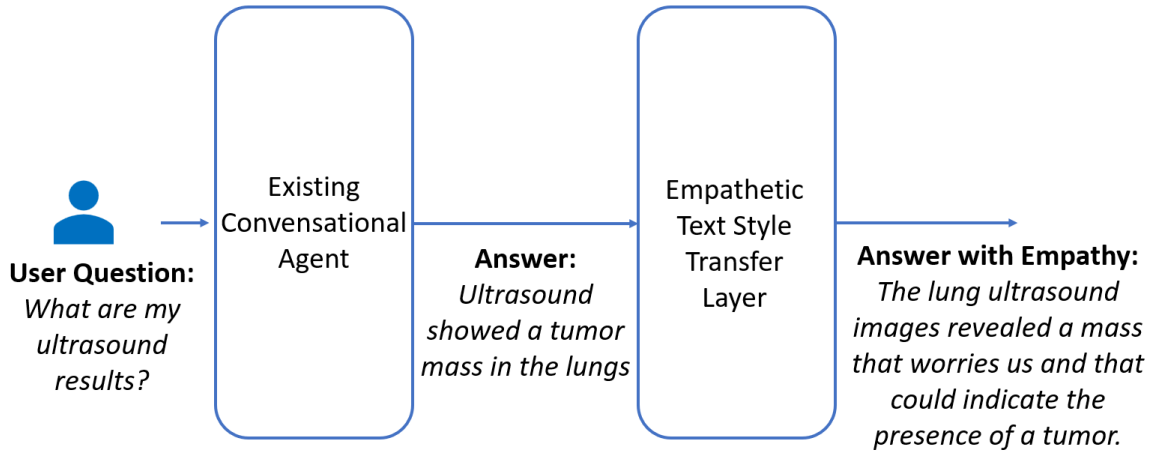

Figure 1.1: Integration of the Empathetic Text Style Transfer layer within an existing conversational agent architecture.

Crucially, this approach enables seamless integration of empathetic capabilities into existing systems without requiring structural modifications or retraining of the base model. For example, a financial chatbot or conversational agent already optimized for regulatory compliance and investment advice can be augmented with empathetic responses simply by cascading the empathetic style transfer layer, thereby improving user experience without compromising precision. Similarly, in healthcare applications,

where clarity and correctness are paramount, the layer ensures that empathetic phrasing does not alter the medical intent of the original message. This modularity makes the solution highly scalable and adaptable across diverse sectors, offering a strategic pathway to human-centric AI interaction while significantly reducing development time, infrastructure costs, and operational risks.

This architecture is particularly advantageous in domains where emotional sensitivity is critical. In mental health support, it facilitates reassurance and validation; in customer service, it improves perceived helpfulness; and in education, it fosters motivational engagement. The modularity of the system ensures scalability and adaptability across sectors, offering a cost-effective pathway to human-centric AI interaction.

To implement and evaluate this approach, we conducted an extensive experimental campaign involving ten small and medium-sized LLMs, selected for their deployability on single-GPU setups. We explored both few-shot learning (FSL) and fine-tuning (FT) strategies using the IDRE (Italian Dialogue for Empathetic Responses) dataset [44], which comprises 480 curated triplets focused primarily on healthcare scenarios. Each triplet includes a user message, an initial chatbot reply, and an empathetically rephrased version.

Model performance was assessed through a comprehensive, multi-layered evaluation strategy. Central to our approach is the adoption of the LLM-as-a-judge paradigm, specifically leveraging the G-Eval framework, which utilizes a state-of-the-art large language model (GPT-4o [30]) as an objective evaluator for the outputs generated by smaller models. This automated evaluation is inspired by established protocols such as SPIKES, widely used in medical communication, and is tailored to ensure a nuanced and contextually relevant assessment across five key dimensions: Empathy, Knowledge, Coherence, Fluency, and Lexical Variety. While the primary focus of the dataset and evaluation was the healthcare domain, this study also systematically explored the generalizability of the style transfer capability by generating and evaluating empathetic responses in additional domains, including work, social, legal, and financial contexts. We validated the G-Eval results and ensured inter-method agreement by integrating a human annotation study. Expert annotators, trained in conversational system evaluation, assessed a representative sample of model outputs using a Likert scale across the same five dimensions; inter-annotator consistency was quantified via Fleiss' kappa. Finally, to fully triangulate and quantitatively support our findings, we integrated traditional NLP similarity metrics BLEU [50], ROUGE [39], and BERTScore [78]. These metrics provided an essential, independent measure of the two core objectives of the task: style transfer (introducing empathetic language) and meaning preservation (maintaining se-

mantic fidelity to the original response). This multi-pronged evaluation framework ensures both the robustness and the interpretability of our results.

The results indicate that fine-tuning large language models with the IDRE dataset is an effective strategy for instilling empathetic capabilities, with particularly promising outcomes on smaller models such as Llama-3.2-1B-Instruct. Moreover, fine-tuning corrected limitations observed in few-shot learning, enhancing both lexical diversity and empathy, as shown in the case of Gemma-3-1B. Finally, the empathetic skills acquired in the medical domain were shown to be transferable to other contexts, suggesting the broad generalizability of empathetic dialogue systems across diverse sectors.

These insights open new opportunities for deploying empathetic AI in a wide range of real-world applications. One potential application is in telemedicine platforms, where empathetic responses can improve patient trust and adherence to medical advice. In elderly care, empathetic conversational agents can provide companionship and emotional support, helping to reduce feelings of isolation. In educational settings, tutors enhanced with empathetic capabilities can adapt their tone to better support students facing difficulties, thereby fostering a more inclusive and motivating learning environment. In addition, legal and financial advisory bots can benefit from empathetic style transfer to deliver sensitive information, such as debt management or legal obligations, in a more considerate and human-centric manner. Finally, in human resources and workplace wellness tools, empathetic AI can assist in managing employee concerns, offering support during stressful periods, and promoting a healthier organizational climate.

Although this research focuses on Italian, the methodology for dataset construction, fine-tuning, and evaluation is intrinsically language-agnostic and can be generalized to other linguistic contexts.

The contributions are twofold. First, the primary innovation lies in the development and validation of IDRE, a resource that addresses a gap in the Italian scientific literature. Second, leveraging this dataset, we propose a novel modular architecture that demonstrates practical and scalable applicability. This architecture incorporates an empathetic reformulation layer within a decoupled design framework, enabling the integration of emotional capabilities into existing systems without requiring costly retraining or structural modifications. This approach effectively addresses both practical and economic challenges associated with empathy integration, significantly reducing development time and infrastructure costs while ensuring adaptability across diverse application domains.

1.3 Structure of the Thesis

The thesis is organized into six chapters, each addressing a specific step of the research pathway, from the conceptual framing of empathy to the practical evaluation of the proposed modular architecture.

Chapter 1 – Introduction

This chapter presents the theoretical foundations of empathy from neuroscientific, psychological, and clinical perspectives, illustrating its relevance within human–computer interaction and, in particular, conversational systems operating in sensitive domains. It outlines the limitations of existing coversational agents or chatbots, the challenges associated with embedding empathetic capabilities in vertical systems, and the rationale for adopting a modular empathetic style transfer layer to enhance emotional expressiveness while preserving factual accuracy.

Chapter 2 – State of the Art

This chapter surveys the principal research contributions on empathetic conversational agents, with specific emphasis on linguistic resources, computational models, and evaluation frameworks. The discussion highlights the substantial lack of Italian parallel datasets for empathetic style transfer and examines empathy as a Text Style Transfer task. The chapter concludes by identifying the methodological gaps addressed by the present work.

Chapter 3 – The Problem

Chapter 3 formalizes the research problem by detailing the scarcity of Italian empathetic corpora, the operational constraints inherent in modifying domain-specific coversational agents, and the associated risks of performance regression. It frames empathetic rephrasing as a constrained Text Style Transfer problem governed by semantic invariance, affective enrichment, and architectural decoupling. The motivations and research objectives are then articulated, providing the conceptual grounding for the methodological approach.

Chapter 4 – The Proposed Approach

This chapter describes the methodological framework adopted in the thesis. It first illustrates the preliminary experimentation based on lexicon-driven emotional masking and discusses the limitations that motivated the shift toward LLM-based dataset construction. Subsequently, it details the creation and human validation of the IDRE dataset and presents the full experimental pipeline, including model selection, fine-tuning strategies, computational setup, and the multidimensional evaluation framework integrating both automated and human assessments.

Chapter 5 – Experimental Results

This chapter reports the experimental results obtained from the evaluation of ten small and medium-sized language models under base, few-shot, and fine-tuned configurations. It examines the performance of the models in the medical domain, assesses their ability to generalize to non-medical contexts, and discusses the alignment between automated and human evaluators. The chapter also analyzes computational efficiency and cost, highlighting the practical advantages of the modular empathetic style transfer layer.

Chapter 6 – Conclusions

The concluding chapter synthesizes the principal contributions of the thesis, including the development of the IDRE dataset, the design of the modular empathetic empathetic style transfer architecture, and the empirical validation across multiple models and domains. It highlights the implications for scalable, human-centric conversational AI and outlines possible directions for future research, such as multilingual expansion, multimodal empathy modeling, and advanced evaluation methodologies.

Finally, to ensure transparency and reproducibility, all resources developed in this study—including the IDRE dataset, generation and evaluation prompts, fine-tuned models, and Python scripts—are publicly accessible through the repositories reported in the appendices. Specifically, Appendix A presents the prompts used throughout the study, Appendix B provides the full Italian versions of the example sentences, Appendix C lists the Hugging Face repositories containing the fine-tuned models, while Appendices D and E report respectively the complete evaluation tables and the full set of graphical visualizations.

Chapter 2

State of the Art

The development of empathetic conversational agents, systems capable of interpreting human affective states and producing contextually appropriate responses, represents a significant shift within Human–Computer Interaction. While early chatbots focused primarily on task efficiency, contemporary research increasingly frames empathy as a functional requirement for user engagement, trust, and safety, particularly in sensitive settings such as mental health, social support, and customer care [8, 74]. This chapter reviews the main lines of research and industry practices in the field, highlighting both international advances and the specific challenges faced in the Italian linguistic and technological ecosystem.

A central asymmetry characterizing current research is the disparity in linguistic resources across languages. In English, the release of Empathetic Dialogues [59] marked a major advancement: a dataset of roughly 25,000 conversations in which a "Speaker" describes an emotionally grounded situation and a "Listener" responds empathetically. Unlike earlier corpora based on emotion labels or keyword polarity, this resource embeds emotional meaning within contextualized narratives, enabling models to learn empathy as a relational and pragmatic behavior. Further developments include datasets such as Counseling and Psychotherapy Transcripts [5], ESConv [41], and DEMO [21], which support multi-turn emotional support, fine-grained empathy annotation, and exploration of specific empathic strategies.

By contrast, the Italian landscape is characterized by a marked scarcity of parallel corpora suitable for generative modeling. Existing Italian resources primarily support classification tasks (e.g., emotion or sentiment recognition) rather than text generation, and often rely on lexicon-based labels rather than context-sensitive annotations [64]. As noted by Mohammad [46], lexicon-driven approaches fail to capture the implicit,

context-dependent nature of affective meaning. Consequently, Italian research faces structural limitations: the absence of datasets comparable to Empathetic Dialogues forces researchers to resort to translation-based methodologies, synthetic augmentation, or rigid rule-based systems, often compromising linguistic naturalness and cultural appropriateness.

Despite these limitations in data resources, the Italian industrial ecosystem has increasingly recognized the strategic relevance of emotional intelligence in automated systems. In mental health applications, platforms such as Unobravo and Serenis have embraced a "Human-First" vision in which AI is employed primarily for tasks of cognitive empathy, such as user profiling and matching with therapists [70, 61, 69]. Fully automated conversational support in Italy—represented by solutions like Wysa [73]—still relies on structured Cognitive Behavioral Therapy scripts. Although users may perceive these interactions as supportive, their rigidity becomes apparent when handling inputs outside predefined domains, leading to the phenomenon of "cheap talk," where responses fail to validate the user's emotional experience [35]. This evidences a broader dichotomy: human-delivered empathy remains accurate but non-scalable, while automated empathy is scalable yet often shallow.

In the corporate sector, Italian companies such as Indigo.ai and Almawave have advanced the integration of affective computing and semantic analysis within customer experience systems. Through sentiment detection, frustration analysis, and adaptive tone-of-voice modulation, these platforms operationalize empathy as clarity, personalization, and accessibility [32, 3]. Yet these systems commonly rely on proprietary, retrieval-based or hybrid architectures that privilege robustness and safety over generative flexibility [33]. Advances in multimodal empathy, such as the development of animated avatars capable of communicating in Italian Sign Language with emotion-sensitive facial expressions [54], represent further progress, although the underlying text generation remains constrained by conventional language models.

Beyond practical constraints, ethical concerns significantly shape research trajectories in this field. Institutions such as the Mario Negri Institute highlight the risks associated with anthropomorphism in companion bots like Replika, including emotional dependency and the illusion of mutuality [35, 34]. These risks echo broader debates on emotional artificial intelligence and its potential to manipulate or inadvertently harm users [14, 65]. As a result, both academia and industry increasingly emphasize transparency, control, and interpretability in empathetic AI systems. From a methodological perspective, the evolution of empathetic generation reflects a transition from explicit lexical manipulation to latent-space control. Early work, such as the Dynamic Emo-

tional Session Generation model [25], enhanced Seq2Seq architectures with dictionary-based attention, treating empathy largely as a lexical augmentation problem. However, the limitations of this approach led to research on disentangled representations, most notably through the Controlled VAE framework introduced by Hu et al. [29], which allows attributes such as sentiment or politeness to be manipulated while preserving semantic content. Subsequent contributions include style transfer models based on cycle consistency [62], adversarial disentanglement [37], and plug-and-play controllers for attribute conditioning [15]. Meanwhile, systematic evaluations of transfer metrics [22, 9] have emphasized the importance of measuring both stylistic accuracy and semantic fidelity, although cross-linguistic inconsistencies persist, particularly for under-resourced languages.

In recent years, the advent of Large Language Models has transformed the field. Models such as GPT 4 and GPT 4.1 demonstrate strong zero-shot performance in empathy-related tasks [72], and studies such as Xu et al. [74] provide multi-dimensional frameworks for evaluating cognitive, emotional, and behavioral aspects of empathetic responses. Nevertheless, LLMs exhibit well-documented limitations, including hallucinations, lack of transparency, inconsistency across domains, and difficulties in satisfying strict evaluative metrics. Furthermore, issues related to data governance, deployment cost, and compliance with regulations such as the EU's GDPR and the AI Act present additional barriers to adoption, especially within European contexts and medium-sized enterprises.

Taken together, academic research and industrial practice reveal a diverse ecosystem of approaches to empathetic conversational modeling. These approaches span:

- rule-based strategies, lexicon-driven augmentation, and scripted behavioral frameworks
- neural architectures based on attention mechanisms, latent disentanglement, and attribute-controlled generation
- evaluation frameworks grounded in style transfer metrics and multidimensional empathy assessments
- applied pipelines combining sentiment analysis, tone adaptation, retrieval-based generation, and multimodal affect recognition

Across this spectrum, high-control but low-flexibility systems coexist with powerful but difficult-to-govern generative models. The field collectively highlights the need

for methods that reconcile semantic stability, stylistic adaptability, transparency, and cross-linguistic applicability—especially for languages such as Italian, where resource scarcity remains a central obstacle to progress.

Chapter 3

The Problem

Despite the remarkable progress achieved in recent years of Large Language Models, most chatbots or coversational agents currently deployed in vertical domains operate under purely informative paradigms. While effective at information retrieval, these systems lack the ability to recognize, validate, and appropriately respond to the user's emotional states, creating a communicative barrier in sensitive contexts.

In the current research and industrial landscape, this limitation is driven by three fundamental unresolved issues that prevent the widespread adoption of empathetic conversational agents, particularly for non-English languages:

- **Linguistic Resource Asymmetry (The Italian Case):** While the English language benefits from structured conversational corpora (e.g., *EmpatheticDialogues*), the landscape for Italian is fragmented. Existing resources are predominantly limited to *Sentiment Analysis* (polarity classification) or single-word labeling. There is a marked absence of public resources providing parallel examples of "neutral" and "empathetic" responses necessary to train supervised natural language generation models.

- **Performance vs. Development Cost Trade-off:** The *end-to-end* development of new empathetic agents is resource-intensive. It requires the collection of vast datasets, expert human annotation, and computationally expensive training cycles. Solutions based on commercial LLMs (e.g., GPT-4) offer "out-of-the-box" empathetic capabilities but introduce operational costs (token/cost) and data privacy issues that are often unsustainable for SMEs or public administrations.

- **Operational Constraints and Regression Risks:** Many organizations utilize established systems optimized over years for specific tasks (e.g., medical triage,

legal consulting). Modifying the "core" of these models to infuse empathy carries a high risk of *catastrophic forgetting* or degradation of factual accuracy. It is operationally impractical to decommission a functioning vertical system to replace it with a less controllable generalist generative model.

Addressing these limitations is not merely a stylistic exercise but a functional necessity, especially in high-emotional-impact contexts such as healthcare. Scientific evidence demonstrates that empathetic communication is not just "pleasant" but concretely improves clinical outcomes: it increases patient treatment adherence, reduces anxiety, and encourages the disclosure of critical information. From an Artificial Intelligence perspective, bridging the gap toward affectivity allows for a transition from *task-oriented* systems to *human-centric* systems, capable of establishing a relationship of trust and emotional safety with the user, reducing the perception of the machine as a cold and detached entity.

To formalize this approach, the problem addressed in this thesis is configured as a constrained *Text Style Transfer* task. To ensure conceptual clarity and avoid philosophical ambiguity regarding machine consciousness, it is fundamental to define the specific scope of "empathy" adopted in this computational framework. We explicitly distinguish between deep cognitive empathy, the biological capacity to fully understand and share another's mental state, which remains beyond the reach of current AI, and surface-level linguistic empathy. This work focuses exclusively on the latter: the operational generation of affective markers and stylistic patterns that elicit a perception of support in the user, without implying any genuine emotional understanding or internal state by the model.

Operationally, the objective is to transform the sterile and purely informative output generated by a traditional chatbot into a response that conveys simulated human warmth, validation, and support, without minimally altering its factual substance. This transformation process is governed by three inseparable fundamental principles:

Semantic Invariance: The most critical requirement is the absolute preservation of the original meaning. The rephrased response must convey exactly the same information as the starting neutral response. In sensitive domains such as healthcare or law, the introduction of empathy cannot afford to modify, omit, or invent facts (e.g., a diagnosis or a bureaucratic procedure), ensuring that the factual truth remains unaltered.

Linguistic Affective Enrichment: Simultaneously with the preservation of meaning, a tangible shift in the communicative register must occur. The system

acts as a stylistic layer, recognizing the implicit emotional context and inserting surface-level linguistic markers of empathy, validation, and kindness (e.g., "I understand this is difficult") that were absent in the original response. The goal is to maximize the user's perception of emotional support through appropriate lexical choices, rather than to replicate human cognitive processes.

Architectural Decoupling: Finally, the proposed solution is defined by its modularity. The empathetic style transfer mechanism must act as an independent layer, positioned "downstream" of the existing chatbot. This means that the empathy system must not require access to the internal workings or training data of the original chatbot but must function as a pure processor of the output text, thereby ensuring maximum compatibility with pre-existing corporate infrastructures.

Ultimately, this work aims to resolve the identified criticalities through two main research directions:

- **Resource Construction (IDRE):** Bridging the linguistic gap through the creation, validation, and public release of the *Italian Dataset for Rephrasing with Empathy* (IDRE). This corpus of 480 triplets (query, neutral response, empathetic response) in the healthcare domain provides the first *ground truth* for the empathetic style transfer task in Italian.
- **Validation of Modular Architecture:** Demonstrating the effectiveness of an "Empathetic Text Style Transfer." Through *fine-tuning* techniques on Small Language Models (SLMs) deployable on a single GPU, we aim to prove that it is possible to infuse affective capabilities efficiently, maintaining informational integrity and drastically reducing costs compared to large-scale generalist LLMs.

Chapter 4

The Proposed Approach

This chapter delineates the methodological framework developed to address the scarcity of empathetic resources for the Italian language. It commences by documenting an initial experimental phase based on lexicon-based sentence corruption, conducted prior to the widespread adoption of LLMs. The analysis of the critical limitations encountered in this antecedent approach, conjoined with the subsequent advent of generative AI, necessitated a strategic paradigm shift. Accordingly, the chapter details the creation and human validation of the IDRE (Italian Dialogue for Empathetic Responses) dataset, a novel resource developed by leveraging the generative capabilities of LLMs to train models in empathetic style transfer task. Finally, the experimental setup is presented, describing the fine-tuning procedures for various architectures and the multidimensional evaluation pipeline adopted to assess their performance.

4.1 Generation Experiment through Lexical Masking

Prior to the widespread adoption of Large Language Models, the investigation focused on the core architectural component: Emotion Rephrasing. This task utilizes generative language models, such as T5 [57] or GPT-2 [56], to perform sentence rephrasing by injecting or amplifying emotional content, relying on supervised training on paired sentence datasets.

The dataset should consist of pairs of sentences, where the first sentence should not express any emotion or, at the very least, have a low emotional content. In contrast, the second sentence should convey the same meaning as the first but express an emotion. Table 4.1 illustrates two pairs of example sentences.

After extensive literature research, it was found that structured datasets for the

Table 4.1: Example of pairs sentences. The sentences presented are English translations of original Italian utterances. The corresponding Italian versions are provided in Appendix B Table B.1.

Neutral Sentence	**Emotional Sentence**	**Emotion**
I can't find the documents I need on the company intranet.	I'm desperate because I can't find the documents I urgently need on the company intranet.	Sadness
We can't guarantee that your hardware issue will be resolved.	We can't promise to fix your hardware issue with certainty, but we'll do our best to help.	Trust

Italian language with such characteristics do not exist. Therefore, one possible approach is to attempt to create it automatically. The proposed idea [13] is to take a dataset of labeled sentences with one or more emotions, such as MultiEmotions-It [64], and create a "corrupted" sentence by masking the emotional content using a labeled lexicon like the NRC-Emotion-Lexicon [46]. The language model's task is then to generate a new sentence with the objective of incorporating an emotion while preserving the semantic meaning. Figure 4.1 illustrates the mechanism.

Figure 4.1: How to feed pairs of emotional sentences/corrupted sentences to a language model. Highlighted in blue are the candidate words to express the emotion of the sentence.

4.1.1 Source Dataset Characteristics (MultiEmotions-It)

For MultiEmotions-It dataset, user comments were extracted from YouTube videos and Facebook ads, resulting in a total of 3240 sentences. These sentences were then annotated by 36 individuals with the basic emotions defined by the Plutchik [51] model (joy, sadness, fear, anger, trust, disgust, surprise, anticipation). It is worth noting that a sentence can express multiple emotions. Each emotion is labeled with a value of 0 if it is not present in the sentence or 1 if it is present. In Table 4.2 are shown some examples of sentences from the dataset:

Table 4.2: Example of pairs sentences in MultiEmotions-It Dataset. The sentences presented are English translations of original Italian utterances. The corresponding Italian versions are provided in Appendix B Table B.2.

Emotional Sentence	**Emotions**	**Emotions Vector**
This song is beautiful and meaningful, great singer♡♡♡♡.	Joy,Trust	1,1,0,0,0,0,0,0
I hate this advert, when it comes on TV it embarrasses me and is hateful.	Anger	0,0,0,1,0,0,0,0
The song is beautiful, it would work well on the radio, unfortunately yesterday at Sanremo Elettra was very emotional, but that's okay, she's human. The music video is high quality, well done, congratulations.	Trust, Sadness	0,1,1,0,0,0,0,0

4.1.2 Critical Issues in Sentence Processing

During the initial analysis of the dataset, several issues were identified, including:

- Sentences composed of few words: This poses a problem during the elimination of lexicon words since there is a risk of completely losing the meaning of the sentence.
- Incomprehensible sentence structures: Some sentences were written in an unintelligible manner, with grammar errors and the use of slang.
- Incorrectly labeled sentences: There were instances where sentences were inaccurately labeled, indicating potential errors made by the evaluators.

4.1.3 Lexicon characteristics

The lexicon consists of 15,256 words, each labeled with the eight basic emotions defined by the Plutchik model. The lexicon was initially created in English and subsequently translated into multiple languages, including Italian. For each emotion, a numerical value between 0 and 1 is assigned, representing its intensity. This intensity value provides a measure of the emotional strength associated with each word in the lexicon.

4.1.4 Limitations and Ambiguities of Translated Lexicon

Indeed, the lexicon posed some significant issues. Firstly, a considerable number of duplicates were identified, accounting for approximately 18% of the dataset. This duplication occurred due to the translation process from English to Italian. As depicted in Figure 4.2, five English words were all translated to the same Italian word, "abbondante." This lack of differentiation in translation introduces ambiguity and hinders the accurate representation of emotions associated with those specific English words in the Italian lexicon.

Word	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Italian Word
abundant	0	0	0	0	0,516	0	0	0	abbondante
copious	0	0	0	0	0	0	0	0	abbondante
plentiful	0	0	0	0	0	0	0	0	abbondante
profuse	0	0	0	0	0	0	0	0	abbondante
rife	0	0	0	0	0	0	0	0	abbondante

Figure 4.2: Lexicon duplicate examples.

Another issue encountered is the presence of many words that are either not associated with any specific emotion or are associated with all the emotions in the lexicon. These words are either irrelevant for the intended purpose or lack the necessary specificity to capture distinct emotional nuances. By filtering out such cases, a reduced dataset is obtained, which accounts for approximately 62% of the original dataset. This filtering process helps refine the dataset by focusing on words that have clear emotional associations, thereby enhancing the quality and relevance of the lexicon for the intended purpose.

4.1.5 Words extraction

To extract the words that trigger the desired emotions from a sentence, a pipeline of cleaning and normalization steps was created. The objective of this pipeline is to identify the most significant words that contribute to the emotional expression within the sentence. The pipeline consists of various stages of data preprocessing, which may include steps such as:

- Convert text to lower case.
- Remove punctuation.
- Tokenization: Breaking down the sentence into individual words or tokens.
- Stopword Removal: Eliminating common words (e.g., articles, prepositions) that do not carry substantial emotional meaning.
- Lemmatization/Stemming: Reducing words to their base or root form to capture their core emotional essence.

in Table 4.3 example of words extracted from sentences using the lexicon and the cleaning pipeline.

Table 4.3: Example of words extracted from sentences using the lexicon and the cleaning pipeline. The sentences presented are English translations of original Italian utterances. The corresponding Italian versions are provided in Appendix B Table B.3.

Emotional Sentence	**Word extracted**	**Corrupted Sentence**
I'm eagerly awaiting the reopening...	['eagerly', 'awaiting']	I'm [EMO_MARK] [EMO_MARK] the reopening...
Really exciting! An ad that makes us dream of a future with optimism.	['exciting', 'optimism', 'dream']	Really [EMO_MARK]! An ad that makes us [EMO_MARK] of a future with [EMO_MARK]
beautiful video I saw it on TV it moved me wow	['beautiful', 'moved', 'wow']	[EMO_MARK] video I saw it on TV it [EMO_MARK] me [EMO_MARK]

Thanks to the pipeline, the word "awaiting" ("aspetto" in italian, first row of the Table 4.3) is detected even though it is not directly present in the lexicon. This is

because lemmatization is applied, and the word "to wait" ("aspettare" in italian) is part of the lexicon. As observed from the "Corrupted phase" column, it is evident that short sentences lose their entire meaning when masking the words from the lexicon. This highlights a limitation of the approach, as the removal of lexicon words can result in the loss of semantic coherence and the overall understanding of the sentence. Please note that I intentionally included a sentence with a grammatical error in the last line as an example. Throughout the analysis, several grammatical errors have been identified, as mentioned earlier.

4.1.6 Dataset and Lexicon validation

To validate the lexicon and determine whether the extracted words are truly indicative of the emotion in a given sentence, a classification task was conducted. The ground truth labels were the emotion vectors associated with the sentences, while the predictions were the intensity values assigned to the extracted lexicon words.

The metrics considered for evaluation were F1-Score, which varies with the threshold, and the Receiver Operating Characteristic curve (ROC) for each emotion. These metrics provide insights into the performance and effectiveness of the lexicon in capturing and predicting the emotional content of the sentences. By analyzing the F1-Score and the ROC curve, it becomes possible to assess the lexicon's accuracy and determine the optimal threshold for considering the emotional relevance of the extracted words.

Specifically, the optimal threshold represents the value that, when applied to the emotion intensity scores of the lexicon, yields the maximum F1-Score. Each word in the lexicon is associated with a value ranging from 0 to 1, representing the intensity of a specific emotion. By setting the optimal threshold, an emotion is considered to be present only if its corresponding intensity value exceeds this threshold. This process is crucial for filtering out low-confidence predictions and improving the overall classification accuracy.

It is immediately noticeable from Figure 4.3 that the performance across all emotions is quite low. Trust appears to have slightly better performance compared to the other emotions. To further investigate, we can examine the occurrence count of each emotion within the entire dataset, as depicted in Figure 4.4. This visualization provides insights into the distribution of emotions and reveals any potential imbalances that could impact classification performance. The class imbalance might affect the model's ability to accurately learn and predict less represented emotions, resulting in lower performance for those classes.

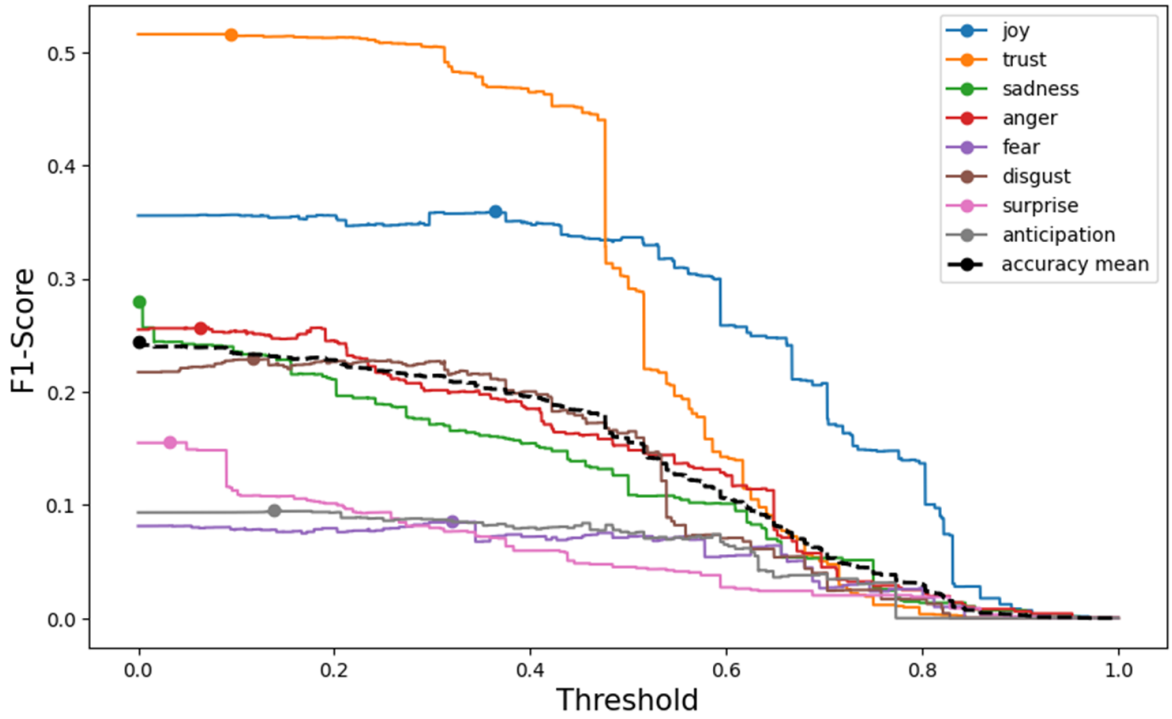

Figure 4.3: F1-Score by varying the Threshold, bullet points represent the optimal value of threshold that maximize the values of F1 for each emotion.

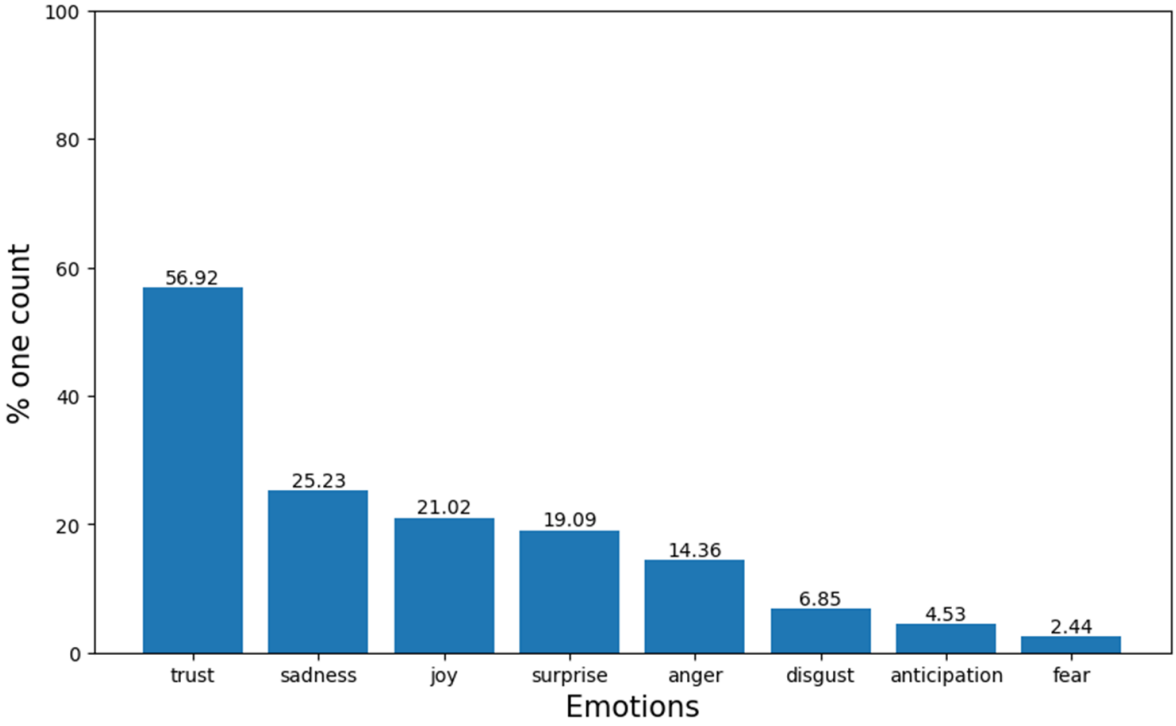

Figure 4.4: Percentage of emotion occurrence in dataset.

Indeed, it is evident that the emotion "trust" has a higher number of occurrences, accounting for 57% of the total dataset, compared to all other classes. On the other hand, the class "fear" is represented by only 2.4% of the sentences. This significant class imbalance indicates a skewed distribution within the dataset, which can contribute to the low performance observed earlier. As shown by the ROC curves in Figure 4.5, the classifier exhibits significantly low performance for each emotion, even when the optimum threshold is set. The results are, in fact, close to those obtainable from a purely random classification.

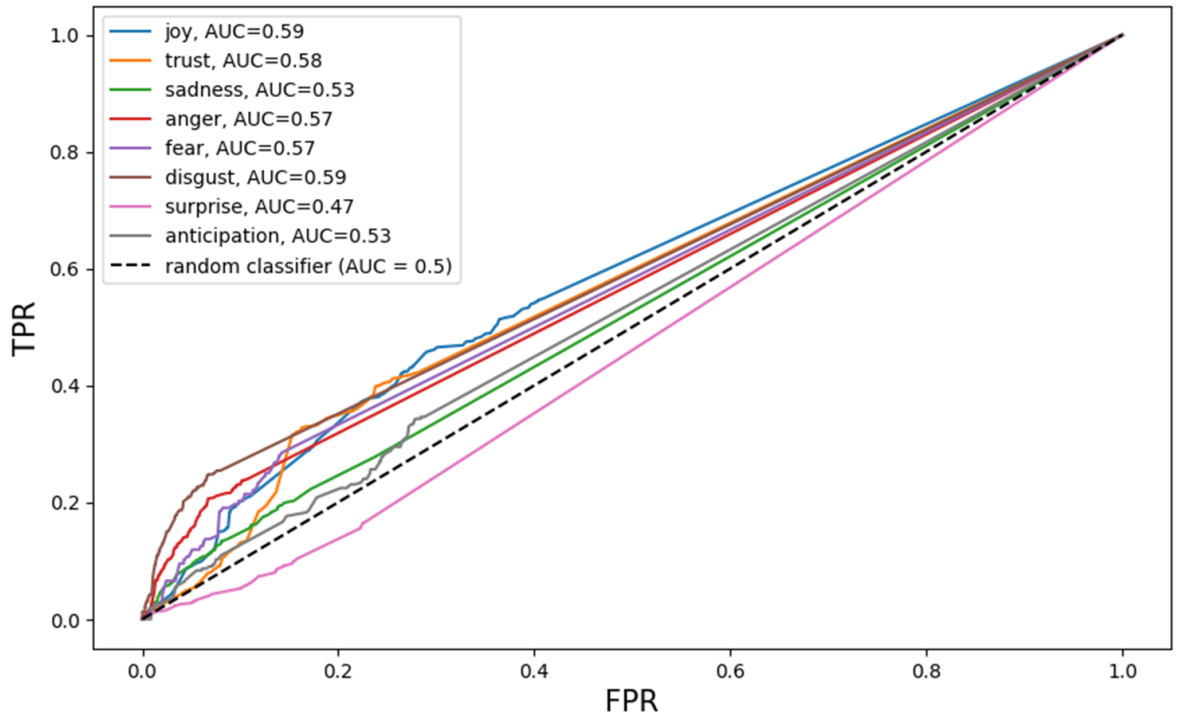

Figure 4.5: ROC curve and Area Under Curve (AUC) for each emotion, FPR is False Positive Rate and TPR is True Positive Rate.

Another issue identified is that during the word extraction phase, short sentences tend to have most of their words eliminated, resulting in their exclusion from the dataset. Considering the combination of these challenges, the decision to abandon this approach and instead attempt to create a new, tailored dataset for this task is understandable. Creating a new dataset specifically designed for the task at hand allows for better control over the data and can address the limitations and imbalances present in the previous dataset. This approach provides an opportunity to improve the performance and reliability of the subsequent analyses and models.

4.2 IDRE Dataset Creation Methodology

The IDRE dataset comprises triplets of sentences, the first sentence represents a user query, the second sentence is the corresponding response generated by a chatbot, and the third sentence is a transformed version of the second sentence intended to enhance its empathetic tone. The sentence generation process was done by the Llama2 13B language model [68], operating on an Azure Virtual Machine equipped with four NVIDIA Tesla V100 GPUs. The choice of Llama 2 was motivated by its open-source nature, which allowed flexible and provider-independent access, and at the time of its creation, represented the state of the art in large language models. The dataset generation process consists of two phases, as illustrated in Figure 4.6:

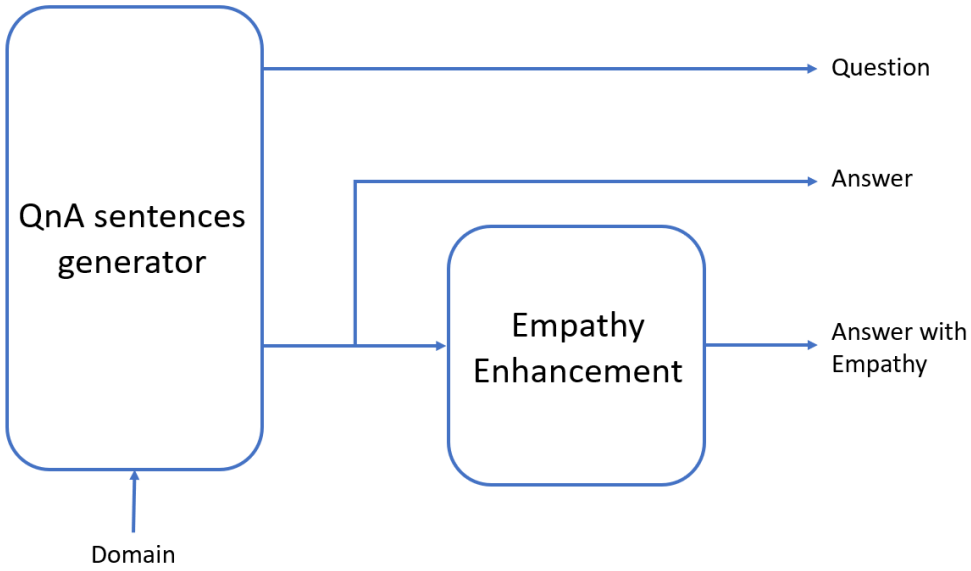

Figure 4.6: Dataset generation process.

QnA Sentence Generation: To ensure the generation of empathetic and compassionate responses, the healthcare domain was selected as the focus for the initial set of bot-human sentence pairs. This domain, characterized by sensitive topics, is well-suited for evaluating the model's ability to generate empathetic responses. The thirteen specific topics chosen for the sentence pairs were invented for the purposes of the experiment:

- information on breast cancer
- breast cancer prevention

- therapies for breast cancer
- psychological support after a cancer diagnosis
- life expectancy after a cancer diagnosis
- psychological support after surgery
- hospital admissions
- post-operative care
- information on leukemia
- psychological support
- anti-cancer therapies
- information on stroke
- preparation for surgeries

An initial set of bot-human sentence pairs was generated using the Llama2 model. These pairs simulated a typical chatbot interaction concerning a specific health issue or domain. For instance, a human query such as "What are the symptoms of COVID-19?" would elicit a corresponding chatbot response like "The most common symptoms of COVID-19 are fever, dry cough, and tiredness".

Empathy Enhancement: After the generation of the initial sentence pairs, an empathy enhancement process was undertaken. Leveraging the Llama2 model once more, the chatbot responses were modified to convey a more empathetic tone. This was achieved by prepending expressions of concern or appreciation, and by substituting specific words to engender a supportive demeanor. To illustrate, the aforementioned chatbot response could be transformed into "I understand that you're concerned about COVID-19. Some common symptoms include fever, dry cough, and fatigue". Both prompts are included in the Appendix A. In Table 4.4 is shown an example of generated sentences.

4.2.1 Human Validation Protocol of the IDRE Dataset

To guarantee the quality of the generated sentences, a rigorous evaluation process was implemented. For each user query, two versions of the chatbot response were evaluated:

Table 4.4: Example of generated sentences. The sentences presented are English translations of original Italian utterances. The corresponding Italian versions are provided in Appendix B Table B.4.

Question	**Answer**	**Answer with empathy**
I heard there are also screenings for breast cancer. What are they?	Yes, there are screenings for breast cancer that can help diagnose the disease at an early stage. The most common screenings are mammography, ultrasound, and biopsy. I recommend talking to your doctor to determine whether screening is necessary for you.	Yes, there are screening tests for breast cancer that can help detect it early. These tests might feel a bit uncomfortable, but they are important for your health. Mammography, ultrasound, and biopsy are the most common screenings, but there are many other methods available. By talking to your doctor, we can determine whether screening is necessary for you and if it's the right time to do it.
What is the average hospital stay for a leg injury?	The length of hospital stay for a leg injury depends on the severity of the injury and the individual patient's condition. On average, the stay can range from a few days to several weeks.	The length of hospital stay for a leg injury can vary greatly depending on the severity of the injury and the patient's individual condition. On average, it can range from a few days to several weeks, but every patient is unique and deserves personalized care and attention for a smooth and speedy recovery.

the standard (non-empathetic) response produced by the original chatbot, and the empathetic response generated by applying the Empathic Text Style Transfer to the same sentence. Twelve volunteer annotators from Lutech Softjam, experienced IT developers and project managers with a solid understanding of the chatbot domain, participated in the study. While their familiarity with the specific domain provided a robust contextual baseline for assessing functional relevance and accelerated the evaluation process, it is acknowledged that the use of non-expert linguists introduces specific methodological considerations. To mitigate potential annotation biases, such as a tendency to prioritize technical accuracy or lexical politeness over nuanced emotional prosody, a structured training protocol was adopted. This involved a calibration session where annotators evaluated a trial set of sentences under supervision to align their subjec-

tive interpretations of "empathy" with standardized linguistic criteria.Each evaluator was assigned 70 sentences: 40 unique to each evaluator for dataset creation, and 30 common to all evaluators solely for measuring inter-annotator agreement with Fleiss' kappa coefficient [20]. Despite the high reliability confirmed by the κ coefficient, the intersection of LLM-generated outputs and non-expert human validation may lead to a degree of "stylistic homogenization." This effect, where both the initial rewriting and the subsequent evaluation converge on statistically probable or conventional linguistic patterns, was monitored to ensure that the IDRE dataset maintains a sufficient variance in emotional expression while remaining bounded by the evaluators' shared mental models of professional chatbot interaction.

Fleiss' kappa was employed to quantify the degree of concordance among multiple annotators, while accounting for agreement occurring by chance. This coefficient ranges from -1 to 1, where negative values indicate agreement below chance, values between 0 and 0.20 suggest slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and values above 0.80 denote almost perfect agreement.

The evaluation was conducted using evaluation dimension-specific questions, requiring responses on a 1 to 5 Likert scale as follows;

1. Totally disagree
2. Disagree
3. Neutral
4. Agree
5. Totally agree

To obtain a more robust analysis and less subject to small variations, the annotation categories were grouped into three macro-categories: scores 1 and 2, score 3 (neutral) and scores 4 and 5. The key evaluation dimensions utilized include:

- **Bot Sentence Correctness:** This evaluation dimension assesses the absence of spelling, grammatical, and punctuation errors in both the user's question and the model-generated response. The evaluation is based on the following criterion:

  "Is the text of the empathic response grammatically and semantically correct?"

- **Absence of English Words in Bot Sentences:** This evaluation dimension verifies whether any English words or phrases are present in the user's question

or the model's response. Exceptions are made for English terms commonly used in Italian (e.g., "badge", "sport"). The evaluation criterion is:

"No English words or phrases are present in the QUESTION and ANSWER columns, unless they are commonly used in Italian."

- **Empathic Answer Correctness:** This evaluation dimension evaluates the absence of spelling, grammatical, and punctuation errors in the model-generated response that includes empathic elements. The evaluation is based on the following criterion:

  "Is the text of the empathic response grammatically and semantically correct?"

- **Absence of English Words in Empathic Sentences:** This evaluation dimension checks for the presence of English words or phrases in the empathic responses generated by the model, excluding those commonly used in Italian. The evaluation criterion is:

  "No English words or phrases are present in the empathic response, unless they are commonly used in Italian."

- **Semantic Coherence:** This evaluation dimension measures the semantic similarity between the standard response and the empathic response generated by the model, ensuring that no concepts are missing or contradictory. The evaluation is based on the following criterion:

  "The empathic response conveys the same semantic meaning as the standard chatbot response. No concepts are missing or contradictory."

- **Empathy Increase:** This evaluation dimension assesses whether the empathic response demonstrates a meaningful increase in empathy compared to the standard response. The evaluation criterion is:

  "The sentence in the ANSWER WITH EMPATHY column expresses more empathy than the sentence in the ANSWER column."

4.3 Experimental Strategy: Fine-Tuning and Setup

The evaluation of the IDRE dataset for empathetic style transfer was conducted by leveraging it both for fine-tuning LLMs and for few-shot learning experiments. The

selection of LLMs followed well-defined criteria aimed at exploring a range of architectures and parameter scales while ensuring computational feasibility. The subsequent sections detail the adopted methodology, including model selection, dataset configuration, computational environment, and training parameters.

4.3.1 Model Selection

In this section ten different LLMs, detailed in Table 4.5, are described, categorized as follows:

- Italian-Optimized Models: Two models were chosen for their specificity or optimization for the Italian language: Minerva-7B-instruct-v1.0 [48] and LLaMAntino-3-ANITA-8B-Inst-DPO-ITA [52]. This selection aims to evaluate the performance of models with a linguistic predisposition tailored to the Italian context.

- Medium-Sized Models (approx. 7-9 billion parameters): Five models fall into this category: Qwen2.5-7B-Instruct [75], Mistral-7B-Instruct-v0.3 [76], Llama-3.1-8B-Instruct [18], granite-3.1-8b-instruct [24] and gemma-2-9b-it [67]. The choice of this parameter range was driven by the necessity of enabling fine-tuning on a single GPU, thereby optimizing computational efficiency.

- Small-Sized Models (approx. 1-3.8 billion parameters): Three smaller models, gemma-3-1b [66], Llama-3.2-1B-Instruct [18], and Phi-3.5-mini-instruct [1], were included to analyze the performance and efficiency of compact models.

This extensive fine-tuning effort seeks to ascertain whether the empathetic transformations within the IDRE dataset effectively translate into enhanced empathetic response generation capabilities across a diverse range of pre-trained models. By systematically assessing the performance of these fine-tuned LLMs, we aim to provide robust evidence regarding the IDRE dataset's contribution to fostering more empathetic and human-centric AI interactions.

4.3.2 IDRE Dataset Preparation

To ensure that only high-quality examples were included in the training set for LLMs, a filtering process was applied to the IDRE dataset. Specifically, all sentences that received a human evaluation score below 3 were excluded. This procedure effectively removed sentences containing linguistic errors or non-contextualized English terms or

Table 4.5: List of language models used in this study, with the corresponding number of parameters and language coverage.

Models	**Parameters [Billions]**	**Language**
Minerva-7B-instruct-v1.0	7.4	Italian
LLaMANTINO-8B-ITA *	8.03	Italian
gemma-2-9b-it	9.24	Multi-Language
Qwen2.5-7B-Instruct	7.62	Multi-Language
Llama-3.1-8B-Instruct	8.03	Multi-Language
Mistral-7B-Instruct-v0.3	7.25	Multi-Language
granite-3.1-8b-instruct	8.17	Multi-Language
Phi-3.5-mini-instruct	3.82	Multi-Language
gemma-3-1b	1	Multi-Language
Llama-3.2-1B-Instruct	1.24	Multi-Language

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

lacking demonstrable improvements in empathy. As a result of this selection, a refined subset of 223 sentences was obtained, each positively rated by evaluators, thus forming a high-quality corpus suitable for model training.

4.3.3 Computational Environment

The fine-tuning of each model was executed on a Microsoft Azure cloud computing platform, utilizing a virtual machine of type `Standard_NC16as_T4_v3`. This configuration was specifically chosen for its processing capabilities, which include 16 vCPU cores, 110 GB of RAM, and 352 GB of disk space. The key computational element for training acceleration was the NVIDIA Tesla T4 GPU, which offers high performance for computationally intensive operations typical of deep neural network training thanks to its Turing architecture and dedicated Tensor Cores. The choice of a single GPU of this caliber allowed for optimization of computational efficiency, enabling the fine-tuning of even medium-sized models (such as those in the 7–9 billion parameter category) within reasonable timeframes.

4.3.4 Training Parameters and Fine-Tuning Strategy

The fine-tuning strategy followed a supervised training approach, where models were exposed to the IDRE dataset to learn to generate empathetic responses. The training parameters were uniformly configured for all ten selected models to ensure a fair and reproducible basis for comparison:

- Learning Rate (LR): A learning rate of 2×10^{-4} was set. This relatively low value is common in fine-tuning pre-trained models to avoid excessive perturbation of already optimized weights and to allow for a finer approximation to the loss function's minimum. The choice of an appropriate LR is critical for balancing convergence speed with training stability.
- Number of Epochs: Each model was trained for 3 full epochs. The number of epochs was empirically determined to balance the model's learning capacity with the risk of overfitting, which refers to the model's tendency to adapt excessively to the training data, losing the ability to generalize to new data. By monitoring performance on the validation set, it was observed that 3 epochs were sufficient to achieve significant improvement without showing clear signs of pronounced overfitting.
- Batch Size: A batch size of 16 was used. The batch size represents the number of training examples processed in parallel before the model's weights are updated. A batch size of 16 is a compromise between computational efficiency (a larger batch size can better utilize GPU resources) and gradient stability (smaller batch sizes can lead to noisier gradients but potentially better local minima). This size allowed for optimized memory utilization of the Tesla T4 GPU while maintaining a stable learning process.
- Random Seed: To ensure reproducibility of the results, the random seed was set to 42.

For the optimization process, the AdamW (Adam with Weight Decay) [43] algorithm was used, a variant of the Adam optimizer that includes weight decay regularization to prevent overfitting. The loss function used was Cross-Entropy Loss, which is standard for sequence generation tasks and language modeling, measuring the difference between the model's predicted probability distribution and the true distribution.

For all models, the application of techniques such as Parameter-Efficient Fine-Tuning (PEFT) [28] was also considered to optimize resources and training times. The key configurations specified are as follows:

- Rank = 16: The rank determines the number of additional parameters that are trained. A value of 16 means that for each weight matrix in the model, two lower-rank matrices are added, one with dimensions (d, 16) and the other with (16, d). These two small matrices replace the fine-tuning of the large original weight matrix, drastically reducing the number of trainable parameters and, consequently, VRAM consumption.

- Target Modules: This indicates the model layers to which LoRA is applied. In this case, the Q (Query), K (Key), V (Value), and O (Output) matrices of the attention mechanism, along with the gate_proj, up_proj, and down_proj layers of the MLP (Multi-Layer Perceptron) block, have been selected for fine-tuning.

- Lora Alpha = 8: This parameter scales the learning of the LoRA matrices. A value lower than r (lora_alpha $<$ r) reduces the importance of the new matrices, acting as a kind of internal learning rate. A value equal to r (lora_alpha = r) is the standard configuration, but here it is set to 8, which implies an implicit regularization that mitigates the risk of overfitting and helps the model generalize better.

- Lora Dropout = 0: Dropout is used to randomly deactivate some neurons during training, preventing overfitting. Setting it to 0 means that this technique is not being used.

- Bias = "none": Bias is an additional parameter added to an output. Setting it to none means no bias is added to the new LoRA layers, which optimizes memory usage.

- Use Gradien Checkpointing = "unsloth": This technique trades computation time for memory, significantly reducing VRAM usage. The unsloth variant is an optimized version that allows training models with very long contexts, reducing VRAM consumption by 30% and enabling larger batch sizes.

4.3.5 Multidimensional Model Evaluation Framework

Recent advances in LLMs have revolutionized automated text generation, introducing unprecedented capabilities [10]. Despite these strides, evaluating the quality of generated text remains a complex challenge. Traditional automatic metrics, such as BLEU, ROUGE, and METEOR [7], primarily based on lexical overlap, often prove inadequate in capturing critical linguistic nuances like semantic similarity, factual consistency, and fluency. While BERTScore represented a significant improvement by incorporating semantic similarity through contextual embeddings, human evaluation persists as the gold standard. This is due to its inherent ability to discern qualitative aspects such as creativity and pragmatic utility; however, it is intrinsically costly, time-consuming, and not scalable for large-scale evaluations [38]. In response to these limitations, the

"LLM-as-a-judge" paradigm has emerged, leveraging the advanced comprehension capabilities of LLMs to assess the quality of generated text. For the implementation of this paradigm, the G-Eval [42] framework was adopted, which comprises the following key components:

- Zero-Shot or Few-Shot Prompting: The "judge" LLM receives instructions via prompts for evaluating the generated text. Such instructions include the input provided to the generator model, the generated output, and, optionally, a reference text.

- Chain-of-Thought (CoT) Prompting: This technique encourages the LLM to articulate its step-by-step reasoning process. This enhances the reliability and transparency of the evaluation, contributing to a reduction in hallucinations and an improvement in judgment consistency [71].

- Criterion-Based Evaluation: The "judge" LLM evaluates the generated text based on predefined qualitative criteria (e.g., fluency, coherence, and relevance). For each criterion, detailed explanations are provided, and a numerical score is assigned, culminating in an overall judgment.

- Scoring Mechanism: The numerical scores assigned by the "judge" LLM are aggregated to derive quantitative metrics, while the rationales generated via CoT support qualitative analysis.

In the present work, we adopted the "Criterion-Based Evaluation" approach.

To better adapt the evaluation to the context of empathetically conveying information, additional evaluation dimensions were defined based on the principles of the SPIKES protocol [6], a well-established model for communicating difficult news in a medical setting. The phases of the SPIKES protocol are described in the following:

- Setting: This initial phase focuses on creating an appropriate environment for the discussion. It includes choosing a private and comfortable location, ensuring sufficient time without interruptions, allowing the presence of significant others (if desired), and adopting open and welcoming body language. The goal is to establish an atmosphere of safety and trust [6].

- Perception: Before delivering the news, it is crucial to explore the patient's understanding and expectations regarding their clinical situation. Questions like "What do you already know about your illness?" or "What is your main concern?" allow

the physician to assess the patient's level of awareness and tailor communication accordingly, avoiding redundant information or further confusion [55].

- Invitation: After understanding the patient's perception, the physician must explicitly ask how much the patient wishes to know. This phase respects patient autonomy and their capacity to manage information. Some patients may want to know all the details, while others might prefer a more gradual approach or delegate some information to family members [11].

- Knowledge: This is the phase where the news is actually delivered. It is crucial to do so clearly, concisely, and using understandable language, avoiding medical jargon. Information should be provided in small "chunks," allowing the patient to process each piece before moving to the next. It is important to regularly check the patient's understanding, for example, by asking "Have I explained myself clearly?" or "Is there anything unclear to you?" Honesty and transparency are fundamental, alongside maintaining an empathetic and supportive tone [6].

- Empathy: After delivering the news, the patient will likely show an emotional reaction (sadness, anger, denial, or fear). The physician must acknowledge and validate these emotions, offering support and understanding. Phrases like "I understand this news is difficult to accept" or "It's normal to feel this way in such a situation" can help the patient feel understood and less alone. A moment of silence can be very powerful in allowing the patient to process information and emotions [53].

- Strategy and Summary: The final phase focuses on planning next steps and summarizing the information provided. Together with the patient, treatment options, care plans, and available support resources are discussed. Realistic goals are set, and any remaining questions are answered. This phase aims to instill a sense of hope and control, even in the face of a difficult prognosis, and to ensure the patient feels supported in their future journey [6].

While acknowledging the comprehensiveness of the SPIKES protocol for physician–patient communication, attention was focused on its "Knowledge" and "Empathy" phases as central dimensions for evaluation, given their high adaptability to the chatbot interaction context.

Evaluation Methodology

As illustrated in Figure 4.7, the model evaluation process is structured into distinct phases. The procedure started with the generation of a synthetic test set using GPT-4o; a selection of examples can be found in Table 4.6. This approach was necessary due to the limitations of the existing IDRE dataset, which was small and had already been used for fine-tuning the model. Creating a new, independent test set was essential to ensure a robust and unbiased evaluation. Leveraging a powerful LLM like GPT-4o enabled the generation of high-quality and diverse sentences.

These new sentences characteristically mimic a chatbot's typical responses: they are concise, direct, and deliver information without overt empathy. The resulting test set consists of 200 sentences: 100 focus on the medical domain, and the remaining 100 are uniformly distributed across four non-medical domains (financial, legal, social, and work-related). The inclusion of diverse domains aims to assess the test set's effectiveness in eliciting empathetic responses even in non-medical contexts.

Table 4.6: Examples of sentences from the synthetic test set, categorized by topic (English translations). The corresponding Italian versions are provided in Appendix B, Table B.5.

Sentence	**Topic**
You have contracted a serious infectious disease.	Medical
You have a severe brain infection that requires urgent treatment.	Medical
The tumor is malignant and has metastasized to several parts of the body.	Medical
You have an incurable autoimmune disease.	Medical
Your company has declared bankruptcy.	Finance
Your mutual fund has experienced significant losses.	Finance
You have been found guilty of tax fraud.	Legal
You have been fined for violating environmental regulations.	Legal
The poverty rate in your municipality has increased by 20% compared to last year.	Social
Your area has been identified as a high-incidence zone for racial discrimination.	Social
Your position has been cut to improve company efficiency.	Work
Your role has been eliminated due to budget cuts.	Work

Each sentence of the test set was reformulated using three different model configurations: the base model (BM), the few-shot learning model (FSL), and the fine-tuned model (FT). The specific prompts utilized for this reformulation are detailed in Ap-

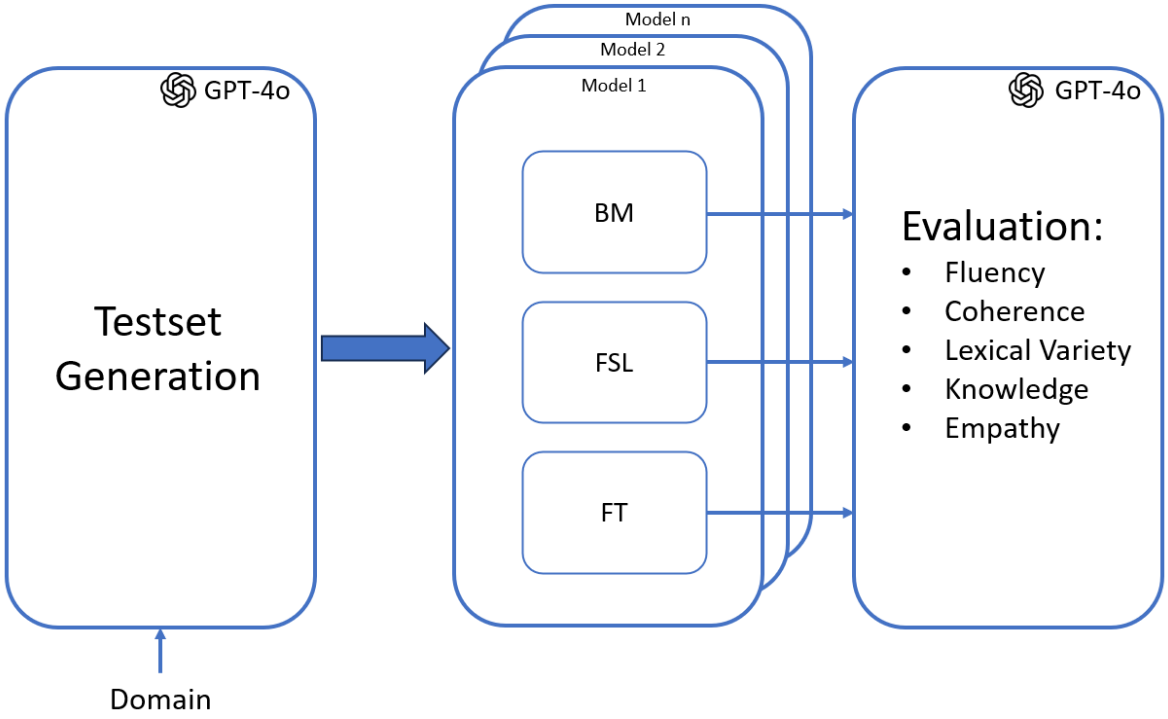

Figure 4.7: Diagram of the evaluation pipeline for empathetic style transfer models, detailing the phases from test set generation to multidimensional assessment.

pendix A. It is important to note that the same core prompt was applied to both the BM and FT models. In contrast, the FSL configuration used an augmented prompt incorporating few-shot examples extracted from the IDRE dataset.

All model outputs were subsequently evaluated using the "LLM-as-a-judge" paradigm across five distinct evaluation dimensions.

- Fluency: Assesses the grammatical correctness, naturalness, and overall readability of the reformulated sentence in Italian.
- Coherence: Measures semantic fidelity, ensuring that the reformulated sentence conveys the same meaning as the original sentence.
- Lexical Variety: Evaluates the diversity of vocabulary used in the reformulated sentence compared to the original.
- Knowledge: Determines how clearly and understandably the reformulated text transmits information, avoiding technical jargon or overly direct phrasing.
- Empathy: Quantifies the extent to which the reformulated text acknowledges and addresses the user's emotional reactions with appropriate empathy.

Each metric was numerically scored on a scale from 1 (lowest) to 5 (highest). While Fluency, Coherence, and Lexical Variety measure the quality of the style transfer itself, the dimensions of knowledge and empathy were specifically derived from the SPIKES protocol to evaluate the communication style.

To validate and corroborate the results obtained through the automatic LLM-based metrics (G-Eval), a human evaluation was integrated. The primary purpose was to verify the inter-methodological agreement between the G-Eval assessments and those conducted by human annotators.

A random subset of 100 sentence pairs was selected from the synthetic test set. Each pair consisted of the original (neutral) sentence and its empathy-augmented version generated by one of the three configurations previously described.

Three independent evaluators, all with proven experience in NLP and chatbots, received a training briefing on the evaluation dimensions and operational instructions. For each sentence, the annotators assessed the response based on the five G-Eval metrics, using a 5-point Likert scale (where 1 = totally disagree; 5 = totally agree). For the purpose of inter-annotator agreement analysis, the scores were aggregated into three macro-categories: [1–2] disagree, [3] neutral, and [4–5] agree. The consistency among evaluators was quantified by calculating Fleiss' kappa coefficient.

In addition to the human assessment, consolidated NLP similarity metrics such as BLEU, ROUGE, and BERTScore were also calculated, which are essential for triangulating the quantitative results. The empathetic style transfer task is inherently dual-objective, requiring both style transfer (shifting from a neutral to an empathetic tone) and meaning preservation (maintaining the factual core). The interpretation of these metrics must therefore reflect this duality.

In our configuration, the original sentence (the neutral source provided to the model) was used as the reference, while the transformed sentence (the empathetic style transfer) was used as the candidate. This setup is crucial for directly measuring both the magnitude of the stylistic transformation and the fidelity to the original message.

The selected metrics are as follows:

- BLEU-4 (Bilingual Evaluation Understudy): This metric was employed to quantify lexical overlap based on n-gram frequency (n = 4). In the context of style transfer, a low BLEU score is expected and paradoxically desirable as it confirms that the model is actively introducing new terminology (e.g., emotional markers) necessary to deviate from the neutral source style.

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Com-

mon Subsequence): This metric assesses the recall of the core content, offering a slightly more permissive view on lexical preservation than BLEU by focusing on the longest common sequence of tokens between the reference and the candidate.

- BERTScore F1: Using contextual embeddings derived from the BERT architecture, this metric measures deep semantic similarity. It serves as the primary indicator for meaning preservation, quantifying the degree to which the factual and conceptual content of the original sentence is maintained in the empathetic transformed sentence, independent of surface-level lexical changes.

Chapter 5

Experimental Results

This chapter presents the experimental results obtained from two complementary analytical components. The first component focuses on the evaluation of the IDRE dataset, examining annotation quality through Fleiss' kappa, analysing score distributions across all linguistic and affective dimensions, and identifying recurrent error patterns. The second component reports the results obtained from the training and evaluation of ten small- and medium-sized Large Language Models. The analysis adopts a multidimensional evaluation strategy to compare Few-Shot Learning and Fine-Tuning , highlighting the decisive role of FT in enabling models, particularly compact architectures, to acquire effective empathetic capabilities. Together, these two components provide a unified view of both the quality of the IDRE dataset and its effectiveness in supporting the development of scalable, linguistically controlled, and empathy-enhanced conversational systems.

5.1 IDRE evaluation results

The evaluation of the IDRE dataset constitutes the first analytical component of this chapter and aims to assess its suitability as a supervised resource for empathetic style transfer in Italian. The analysis focused on three aspects:

- inter-annotator agreement
- distribution of human ratings across linguistic and affective dimensions
- qualitative inspection of recurrent error patterns

The results, summarized in Table 5.1, reveal that the highest agreement levels were observed for evaluation dimensions related to the presence of English words likely due

to the relative simplicity of this annotation task. In contrast, evaluation dimensions involving more nuanced linguistic features yielded lower, yet still acceptable, agreement levels, generally falling within the moderate range.

Table 5.1: Fleiss' Kappa coefficient for each evaluation dimension. The "Aggregated Fleiss' Kappa" aggregates scores into three macro-categories: 1–2, 3 (neutral), and 4–5.

Evaluation dimensions	**Fleiss Kappa**	**Aggregate Fleiss Kappa**
Bot sentence correctness	0.608	0.821
Absence of English words in bot sentences	0.781	0.927
Empathic Answer correctness	0.566	0.807
Absence of English words in empathic sentences	0.587	0.881
Semantic coherence	0.587	0.881
Empathy increase	0.645	0.840

The evaluation of the IDRE dataset demonstrated an overall satisfactory quality, as illustrated in Figure 5.1. Across all assessed evaluation dimensions, the generated sentences consistently received positive ratings from annotators, as evidenced by the predominance of favorable evaluations (light blue bars in the chart).

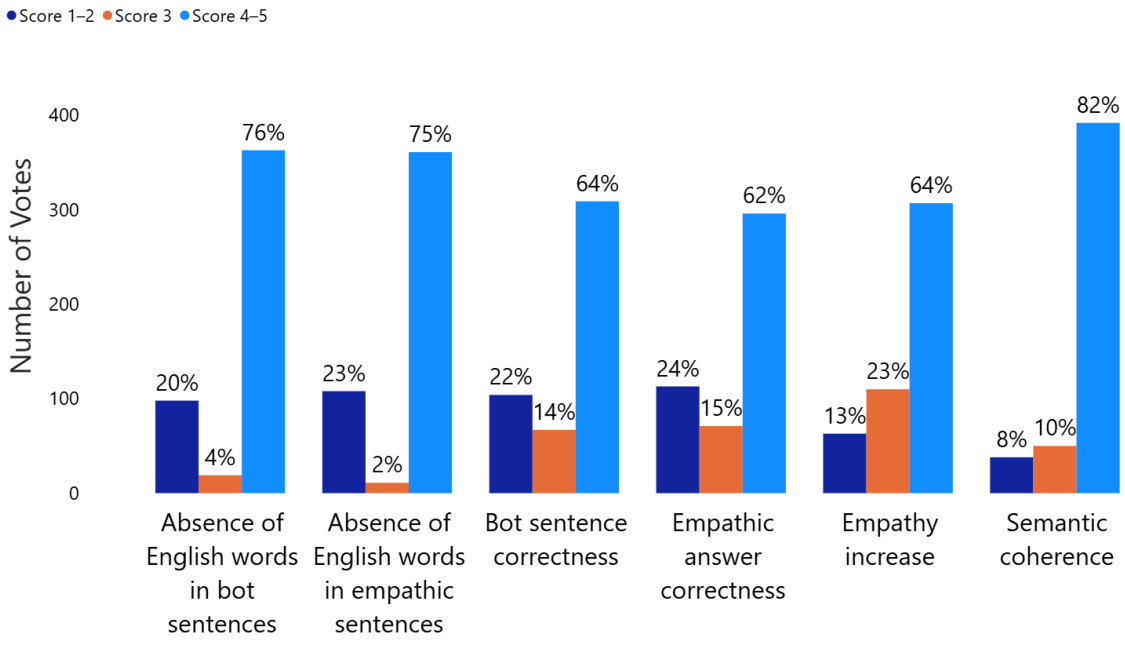

Figure 5.1: Score distribution for each evaluation dimension.

For instance, the Semantic coherence metric received 82% of positive ratings, 10% of neutral, and only 8% of negative ones, indicating a strong adherence of the responses

to the intended meaning. This result suggests that the enhancement of empathetic tone did not compromise the informational quality of the responses. On the contrary, it enriched their communicative dimension, improving emotional resonance while preserving semantic content.

However, a more in-depth analysis of the evaluation dimensions with lower scores highlighted two main areas for attention and improvement: grammatical errors and the presence of non-Italian terms. A substantial subset of responses exhibited recurring grammatical issues. A representative example is the sentence: *"Ohimini, cara/o utente, è comprensibile che durante il trattamento del tumore possa esserti difficile gestire i sintomi. Sono qui per aiutarti a trovare soluzioni e supporti per farcela insieme".*, which contains the non-existent term *"Ohimini"* and a typographical error in the word *"supporti"*. Moreover, several sentences included non-Italian terms, predominantly English, as in *"Per prevenire le infezioni after surgery, è importante seguire le istruzioni del medico e del personale ospedaliero, come ad esempio lavare le mani frequentemente, evitare di toccare la ferita e utilizzare dispositivi di protezione individuali."*. This phenomenon is likely due to the multilingual nature of the language model employed, which was selected because, at the time this work was carried out, a dedicated Italian model was not yet available.

These findings underscore the importance of refining the model training process by prioritizing Italian vocabulary and syntactic structures. Alternatively, employing a model natively developed or specifically fine-tuned for Italian could significantly enhance both linguistic accuracy and expressive naturalness.

Finally, although the increase in empathetic tone was positively evaluated in 306 sentences, a non-negligible portion of the responses (173 sentences, equal to 36% of the total) did not show a significant improvement compared to the original versions. This observation highlights the potential for further optimization of the generation process to ensure a more consistent and meaningful enhancement in affective communication.

5.2 Models evaluation results

The second analytical component of this chapter presents the evaluation results obtained from ten small- and medium-sized LLMs trained using the IDRE dataset. The objective is to measure how effectively these models acquire and apply empathetic style transfer capabilities under different adaptation strategies—namely Few-Shot Learning and Fine-Tuning. The analysis follows a structured approach, articulated into five sub-sections: (i) Medical Domain Performance – assessing the models' behaviour in

the domain represented in the original dataset; (ii) Cross-Domain Generalization – analysing performance on legal, financial, social, and workplace topics, to evaluate stylistic transferability; (iii) Italian Model Evaluation – comparing models specifically optimized for Italian with multilingual counterparts; (iv) Inter-Rater Agreement and Metric Validation – examining how automated evaluation (G-Eval) aligns with human judgment, and supporting the analysis with BLEU, ROUGE, and BERTScore metrics; and (v) Time and Cost Analysis – quantifying the computational efficiency and cost-effectiveness of the modular empathetic style transfer layer.

This multidimensional framework enables a comprehensive comparison of FSL and FT strategies, revealing the conditions under which empathetic style transfer can be reliably integrated into compact, resource-efficient LLMs suitable for industrial deployment.

5.2.1 Medical Domain Performance

The experimental analysis conducted in the medical domain highlights that Empathy is the dimension that benefited most from the model adaptation strategies, emerging as the primary and most significant result of this study. While FSL configurations showed limitations in particular on the small models, FT proved decisive in unlocking the affective capabilities of the models, particularly for more compact architectures.

The most relevant result is the drastic increase in the models' capacity to generate emotionally resonant responses after Fine-Tuning. As highlighted in Figure 5.2, which isolates the progression of the empathy score for the Gemma-3-1B model, an exceptional qualitative leap is observed: the score rises from a critical value of 1.59 in FSL mode to a near-perfect 4.93 in FT mode.

A similar behavior is observed in the Llama-3.2-1B-Instruct model, illustrated in Figure 5.3.

Starting from a baseline (BM) score of 2.71, the model suffers a regression in FSL mode (1.72), only to reach a score of 4.59 thanks to Fine-Tuning. These data confirm that for small-sized models (approximately 1 billion parameters), the FSL approach is insufficient to abstract the complex concept of "empathetic style", often leading to performance regression. Conversely, FT allows these "lightweight" models to internalize the affective patterns of the IDRE dataset, surpassing larger models in effectiveness and achieving percentage increases in empathy greater than 30% compared to baselines (see Table 5.2).

The overall mean increments are reported together with their 95% confidence inter-

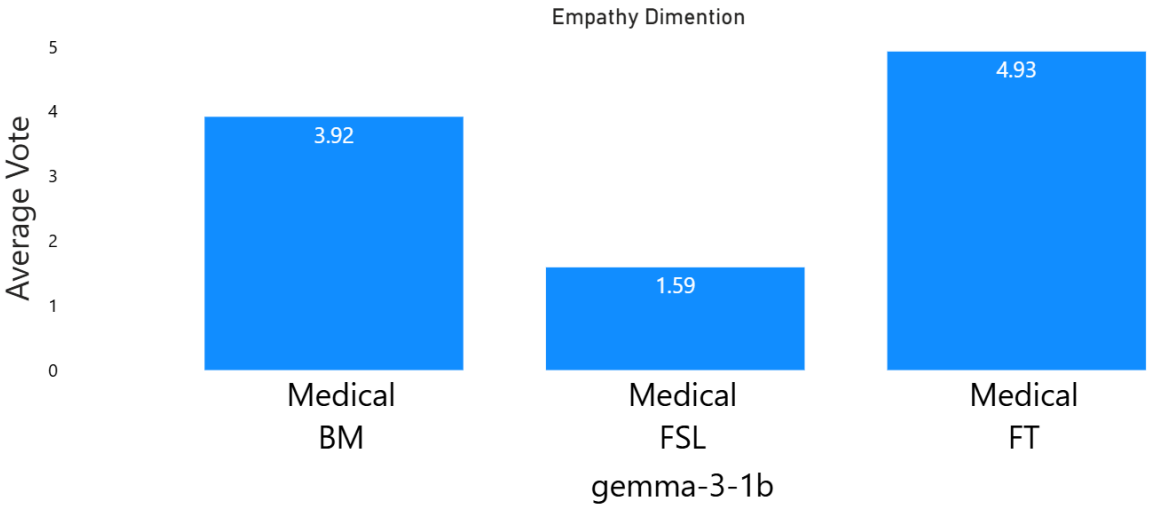

Figure 5.2: Progression of the Empathy score for the Gemma-3-1B model. Note the significant recovery in FT compared to the collapse in FSL.

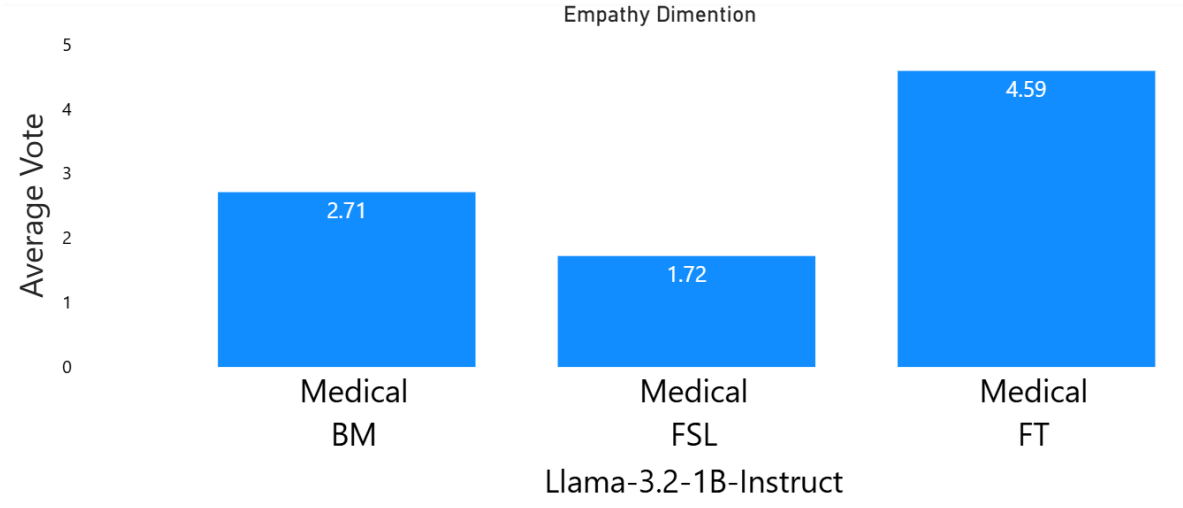

Figure 5.3: Progression of the Empathy score for the Llama-3.2-1B-Instruct model. Fine-Tuning enables high emotional resonance.

Table 5.2: Percentage improvement in evaluation scores for models using FT and FSL models compared to the BM. The IC values are FT: 0.067 (95% CI: 0.001, 0.133) and FSL: 0.013 (95% CI: −0.053, 0.080).

Models	FSL Increment (%)	FT Increment (%)
Llama-3.2-1B-Instruct	17.08	32.34
LLaMANTINO-8B-ITA *	5.06	9.22
gemma-3-1b	−20.42	4.93
Mistral-7B-Instruct-v0.3	3.64	4.21
llama-3.1-8B-Instructt	1.40	3.28
Phi-3.5-mini-instruct	−1.48	3.07
gemma-2-9b-it	−1.92	2.92
Qwen2.5-7B-Instruct	2.98	2.62
Minerva-7B-instruct-v1.0	4.19	2.41
granite-3.1-8b-instruct	3.03	2.31

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

vals (CIs): FSL 0.013 (95% CI: −0.053, 0.080) and FT: 0.067 (95% CI: 0.001, 0.133).

Although empathy represents the main focus, analyzing linguistic stability metrics (*Fluency, Coherence, Lexical Variety*) reveals a critical insight regarding the limitations of Few-Shot Learning. As visualized in Figures 5.4 and 5.5, small models in FSL configuration maintain surprisingly high Coherence scores (e.g., 4.73 for Gemma-3-1B). However, this data presents a 'Coherence Paradox': qualitative analysis confirms that this high score is not due to semantic fidelity during style transfer, but rather to the model's tendency to repeat the input verbatim. This is corroborated by the collapse of *Lexical Variety* (1.28 for Gemma-3-1B in FSL).

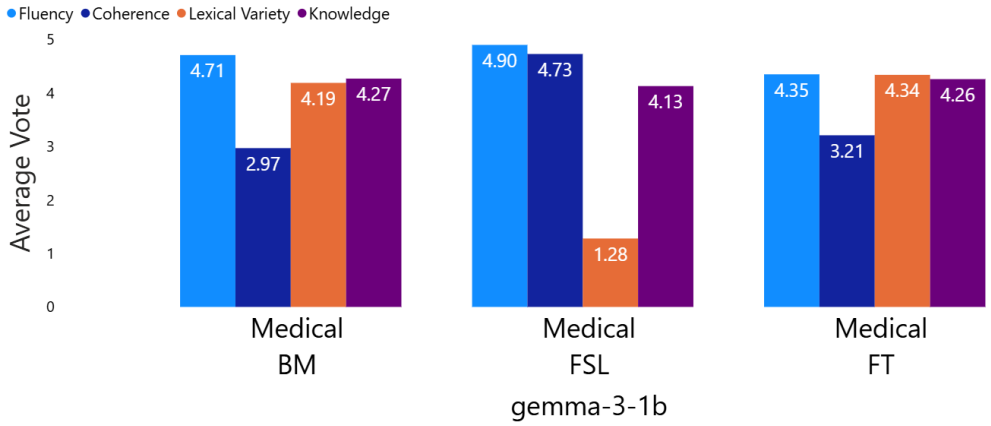

Figure 5.4: Structural metrics for Gemma-3-1B. The FT configuration restores Lexical Variety while maintaining acceptable levels of Coherence.

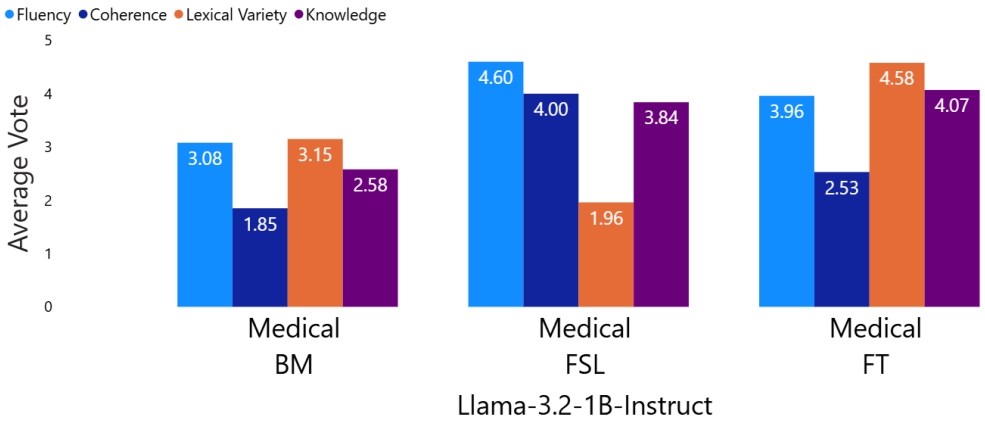

Figure 5.5: Structural metrics for Llama-3.2-1B-Instruct. The chart highlights the stability of Knowledge and Fluency across configurations.

Fine-Tuning effectively resolves this paradox. While it causes a slight physiological dip in Coherence (due to the structural changes required by emotional markers), it restores Lexical Variety to excellent levels (4.34). This confirms that the Fine-Tuned model actively rephrases the content, balancing informational precision with expressive richness.

To illustrate the qualitative impact of these configurations, Table 5.3 presents an example generated by the LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model.

Table 5.3: Examples of sentences for the LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model. The sentences included are English translations of original Italian utterances. The corresponding Italian versions are provided in Appendix B, Table B.6.

Sentence	**Model Configuration**
You have a severe respiratory infection that requires urgent treatment.	Original
You are in a medical emergency due to a very serious respiratory infection that requires immediate medical intervention.	Base Model (BM)
I'm concerned about the severity of your respiratory infection and want to ensure you receive the most appropriate treatment in a timely manner.	Few-Shot Learning (FSL)
Your health condition appears to be concerning; you've been diagnosed with a serious respiratory infection that requires urgent medical intervention to prevent potentially serious consequences.	Fine Tuning (FT)

This comparison highlights a critical safety implication for healthcare applications.

In the FSL configuration, the model prioritized polite expressions ("*I am concerned*") but failed to convey the urgency of the treatment, effectively diluting the medical message. Conversely, the Fine-Tuning process successfully balanced the stylistic requirement of empathy with the constraint of informational completeness. This suggests that Fine-Tuning acts not only as a style transfer mechanism but also as a *semantic safety anchor*, preventing the "toxic positivity" phenomenon where emotional tone overshadows factual risks.

Table 5.4 presents the detailed relative performance improvements of all models across all evaluation dimensions, confirming that Empathy consistently emerges as the metric with the most substantial improvement in the FT configuration.

Table 5.4: Performance improvement (in percent, %) of models in FT and FSL configurations over the baseline model, reported for each evaluation dimension.

Model	**Fluency**		**Coherence**		**Lexical Variety**		**Knowledge**		**Empathy**	
	FSL	**FT**	**FSL**	**FT**	**FSL**	**FT**	**FSL**	**FT**	**FSL**	**FT**
gemma-2-9b-it	0.62	3.61	−51.09	−1.96	7.96	2.89	0.83	0.42	13.18	9.26
gemma-3-1b	3.88	−8.28	37.21	7.48	−227.30	3.46	−3.39	−0.23	−146.50	20.49
granite-3.1-8b-instruct	0.20	2.00	−16.24	−3.68	16.07	6.70	2.40	1.21	10.31	7.05
Llama-3.1-8B-instruct	3.64	4.02	−31.58	2.07	10.35	2.81	7.33	3.19	8.17	4.26
Llama-3.2-1B-instruct	3.30	2.22	53.75	6.88	−60.71	5.22	15.81	16.61	−57.56	40.96
LLaMANTINO-8B-ITA *	4.68	2.30	−50.19	7.11	15.53	5.34	13.23	9.50	20.11	23.74
Minerva-7B-instruct-v1.0	2.63	−6.64	−13.73	4.44	0.75	4.81	14.58	−1.46	10.11	12.73
Mistral-7B-instruct-v0.3	11.47	6.98	−11.62	5.68	7.05	−0.99	1.44	−0.84	4.30	9.42
Phi-3.5-mini-instruct	2.14	2.56	−35.03	1.49	7.32	0.98	4.23	4.67	2.50	5.11
Qwen2.5-7B-instruct	0.83	2.66	−31.44	7.77	20.60	−3.31	5.44	0.21	11.21	4.20

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

5.2.2 Cross-Domain Generalization

An analysis of model performance on sentences from non-medical domains reveals that most models exhibit a satisfactory ability to generalize, maintaining empathy levels comparable to those observed in the medical context. This trend is illustrated in Figure 5.6, which reports the average scores achieved by four models (Qwen2.5-7B-Instruct, Phi-3.5-mini-instruct, Llama-3.2-1B-Instruct, and LLaMAntino-3-ANITA-8B-Inst-DPO-ITA) across all topics. Notably, the models' performance in finance, legal, social, and welfare domains appears broadly aligned with their results in the medical domain.

Table 5.5 provides a comparative analysis of empathy-related performance improvements, measured in both FSL and FT configurations relative to the BM.

However, some exceptions emerge. The Minerva-7B-instruct-v1.0 model, for instance, showed significantly lower performance in non-medical domains, with negative

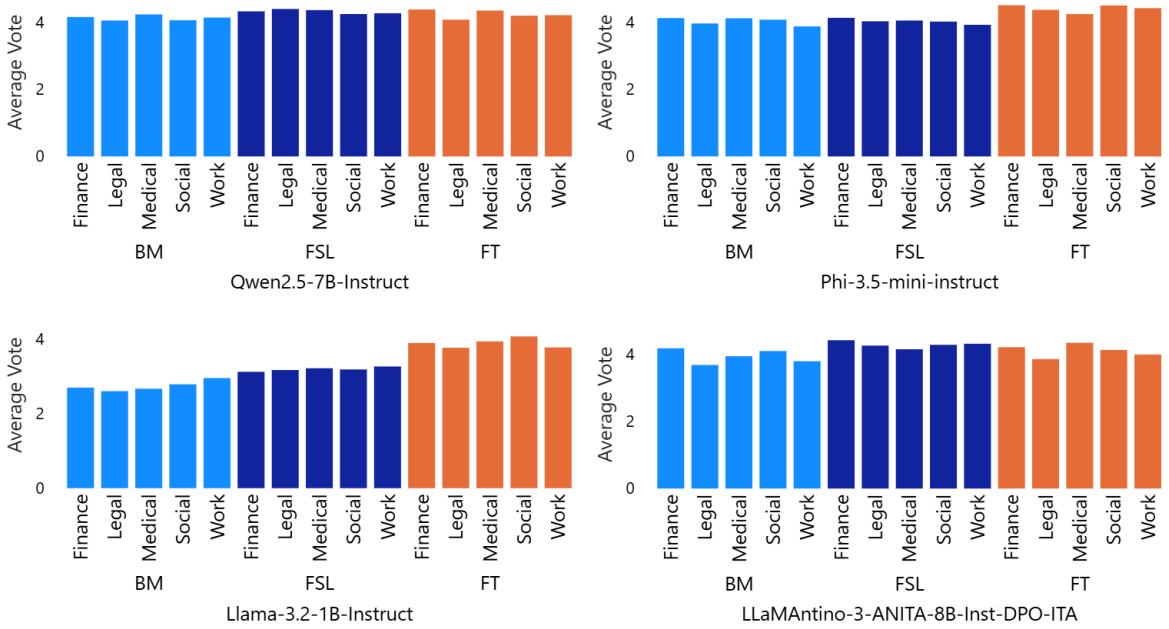

Figure 5.6: Average scores per evaluation dimension for selected models across multiple domains.

empathy improvements in almost all cases (finance: FSL −17.86%, FT −7.61%; legal: FSL −34.57%, FT −45.33%; work: FSL −26.15%, FT −7.89%; and social: FT −9.64%). This contrasts with its positive gains in the medical domain (FSL +10.11%, FT +12.73%), suggesting limited generalization capabilities.

Table 5.5: Empathy-related performance improvements for each model in FT and FSL configurations, reported by domain. All values are expressed as a percentage (%).

Model	**Finance**		**Legal**		**Social**		**Work**		**Medical**	
	FSL	**FT**	**FSL**	**FT**	**FSL**	**FT**	**FSL**	**FT**	**FSL**	**FT**
gemma-2-9b-it	12.61	9.35	27.83	13.54	18.02	17.27	36.21	8.64	13.18	9.26
gemma-3-1b	−72.31	8.20	−112.20	22.32	−253.13	6.61	−71.11	28.70	−146.54	20.49
granite-3.1-8b-instruct	4.10	4.10	12.00	1.79	7.20	7.20	8.20	5.08	10.31	7.05
Llama-3.1-8B-instruct	16.81	12.96	30.36	29.73	13.16	12.39	33.02	11.25	8.17	4.26
Llama-3.2-1B-instruct	−92.11	35.40	−27.08	44.55	−82.86	43.36	−64.71	48.15	−57.56	40.96
LLaMANTINO-8B-ITA *	24.59	−1.10	48.31	8.96	31.90	7.06	52.85	13.43	20.11	23.74
Minerva-7B-instruct-v1.0	−17.86	−7.61	−34.57	−45.33	3.19	−9.64	−26.15	−7.89	10.11	12.73
Mistral-7B-instruct-v0.3	9.09	15.29	14.78	10.91	−2.83	8.40	10.28	9.43	4.30	9.42
Phi-3.5-mini-instruct	9.52	18.10	3.81	15.13	5.77	18.33	2.88	9.82	2.50	5.11
Qwen2.5-7B-instruct	12.26	17.70	18.97	−9.30	−7.22	3.70	14.81	6.12	11.21	4.20

* LLaMANTINO-8B-ITA si riferisce al modello LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

Smaller models, such as gemma-3-1b and Llama-3.2-1B-Instruct, displayed a recurring pattern: suboptimal performance in the FSL configuration, followed by substantial improvements after fine-tuning. This underscores the critical role of FT in adapting lightweight models for empathetic tasks across diverse domains.

Conversely, models like Phi-3.5-mini-instruct and Llama-3.1-8B-Instruct achieved greater empathy improvements in non-medical domains compared to the medical one. For example, Phi-3.5-mini-instruct recorded a peak improvement of 18.33% (FT, social) versus 5.11% in the medical domain, while Llama-3.1-8B-Instruct reached 33.02% (FSL, work) compared to 8.17% (FSL, medical). These results suggest that such models may be inherently more versatile in general contexts than in specialized ones.

Finally, the granite-3.1-8b-instruct model maintained stable performance, with small but consistent empathy improvements across all domains. This behavior likely reflects its high intrinsic empathy baseline, leaving a narrower margin for improvement compared to models starting from lower initial performance.

5.2.3 Italian Model Evaluation

An additional focus of this study is the evaluation of sentence generation quality in models specifically tailored for the Italian language:

- LLaMAntino-3-ANITA-8B-Inst-DPO-ITA demonstrates substantial improvements over the baseline, with performance gains of +5.06% in FSL and +9.22% in FT as shown in Table 5.2. These results highlight the effectiveness of the combined instruction tuning and DPO approach, particularly when supported by deep adaptation via fine-tuning. Notably, LLaMAntino excels in the empathy dimension, achieving a +23.74% improvement in FT one of the highest scores across the entire evaluation (Table 5.5). The model also demonstrates balanced performance across other dimensions: it enhances text fluency and lexical diversity, indicating strong stylistic control and vocabulary enrichment without compromising semantic coherence. However, structural coherence shows some instability in FSL, which is effectively mitigated through FT, suggesting that supervised optimization can address architectural limitations. In terms of knowledge integration, LLaMAntino performs well, with consistent improvements likely attributable to careful source selection during training.

- Minerva-7B-instruct-v1.0 presents a more nuanced profile. As shown in Table 5.2, It achieves greater improvement in FSL (+4.19%) than in FT (+2.41%), suggesting strong intrinsic comprehension and generalization capabilities, likely stemming from its pre-training phase. The model appears particularly well-suited for extracting relevant information from limited examples, reducing reliance on fine-tuning. In the empathy dimension, Minerva also performs well, with gains of

+10.11% in FSL and +12.73% in FT (Table 5.5), although less pronounced than LLaMAntino. In the knowledge dimension, Minerva surprisingly outperforms the baseline in FSL, but exhibits weaknesses in coherence and fluency under FT, indicating a higher sensitivity to adaptation strategies. Lexical diversity improves moderately, and empathetic expression remains one of its core strengths.

5.2.4 Inter-Rater Agreement and Metric Validation

To assess the impact of including G-Eval as an annotator, Fleiss' kappa coefficient was calculated under two scenarios: three human annotators and a combination of two human annotators plus G-Eval. All human annotators were experts in the NLP domain and received dedicated training on the evaluation criteria prior to the assessment. The total number of raters was kept constant to avoid the statistical artifact introduced by increasing the number of annotators, which tends to lower the kappa value simply due to the higher probability of disagreement. This methodological choice enables us to attribute any observed variation in kappa specifically to the inclusion of G-Eval. The results obtained are reported in Table 5.6.

Table 5.6: The table presents Fleiss' kappa coefficients for each evaluation metric, comparing the Human-Only and 2 Humans + G-Eval configurations. The final column reports the difference in agreement between the two scenarios.

Evaluation Dimensions	**Kappa (Only Humans)**	**Kappa (2 Humans + G-Eval)**	**Delta**
Fluency	0.656	0.649	−0.007
Coherence	0.606	0.565	−0.041
Lexical Variety	0.405	0.315	−0.090
Knowledge	0.429	0.405	−0.024
Empathy	0.541	0.449	−0.092

The findings show a reduction in agreement across all metrics, with decreases ranging from −0.007 (fluency) to −0.092 (empathy). Fluency and knowledge exhibit only marginal decreases, suggesting that G-Eval is relatively consistent with human raters when evaluating structural and informational aspects. Coherence and empathy show more pronounced reductions, confirming that semantic and affective dimensions are more challenging for an automated model to replicate. Lexical variety records a significant drop, likely due to divergent interpretations of sentences where the style transfer consisted of adding a brief empathetic segment at the beginning or end of the original

sentence. In these cases, some annotators reported uncertainty about the appropriate score, oscillating between low ratings (since most of the text was unchanged) and high ratings (because the addition introduced a stylistic change perceived as relevant). It is important to note that empathy is inherently subjective: the annotators reported difficulties in establishing uniform criteria, particularly when the empathetic tone was implicit or ambiguous, which contributed to increased variability in the ratings. Similarly, for knowledge, some raters highlighted the complexity of distinguishing between informational clarity and empathetic tone, resulting in variability in the ratings.

In addition to inter-rater agreement, we further assessed the outputs using automatic similarity metrics to provide a complementary perspective on style transfer and semantic preservation, as summarized in Table 5.7.

Table 5.7: Aggregate results of automatic similarity metrics (BLEU-4, ROUGE-L, and BERTScore F1) for the evaluation of style transfer and semantic coherence.

Evaluation Dimensions	**Measured Property**	**Calculated Result (Mean)**
BLEU-4	Lexical Overlap	0.17
ROUGE-L	Content Recall	0.38
BERTScore F1	Contextual Semantic Coherence	0.70

The combination of these results strongly supports the hypothesis that the Empathetic Text Style Transfer operates effectively, achieving its objectives while maintaining appropriate constraints on meaning.

The low scores for BLEU-4 (0.17) and ROUGE-L (0.38) validate the success of the style transfer. These values confirm that the model's output is not a mere verbatim copy or a trivial synonym substitution of the source sentence. Instead, the model introduced significant lexical and structural changes, such as opening phrases for empathy and modal modifiers, which necessarily penalize n-gram-based metrics but are essential for the required emotional shift.

The most critical finding for substantiating the model's viability is the BERTScore F1 value of 0.70. This result demonstrates a substantially preserved level of semantic coherence between the original neutral message and the empathetic output. Although this score is not maximal (which is typically in the 0.85–0.95 range for pure paraphrasing tasks where the style is unchanged), the 0.70 figure must be interpreted within the inherent style transfer vs. meaning preservation trade-off.

The observed deviation from the ideal BERTScore is a direct consequence of the empathetic intervention. Introducing subjective, emotional modifiers (e.g., "I understand

this is difficult," or "I am sorry to hear that") inevitably alters the global semantic vector of the sentence, even when the factual kernel remains untouched. This slight reduction in semantic identity is thus considered the necessary cost for achieving emotional efficacy. The 0.70 value is strong evidence that the model successfully avoids semantic drift or factual inconsistencies, confirming that the empathetic style transfer layer is a controlled mechanism that enhances emotional intelligence without fundamentally compromising the informational integrity of the base task.

5.2.5 Time and Cost Analysis

This section provides a quantitative assessment of the time and costs associated with the fine-tuning process and subsequent phrase generation. As outlined in Section 4.3.1, the evaluated LLMs were grouped into two categories based on their parameter count. All measurements were conducted using a Standard_NC16as_T4_v3 virtual machine within the Microsoft Azure cloud environment, as described in Section 4.3.3. To evaluate efficiency and cost-effectiveness, a benchmark was performed by generating 100k empathetic phrases per model.

The fine-tuning analysis revealed that small-scale models required an average of 5 min per training session, with an estimated cost of EUR 0.08. In contrast, medium-scale models exhibited longer training durations, ranging from 8 to 15 min, with corresponding estimated costs between EUR 0.15 and EUR 0.30.

For the generation task, small-scale models produced 100k phrases in approximately 17 h, incurring a cost of EUR 18. Medium-scale models completed the same task in around 50 h, with an estimated cost of EUR 54. For comparative purposes, the cost of generating 100k equivalent empathetic phrases in Italian via human annotators on the Amazon Mechanical Turk (MTurk) [4] platform was estimated. Assuming a reward of USD 0.1 per phrase and a 20% platform fee, the total cost for human-based generation amounted to USD 12k (approximately EUR 11,280). To provide a benchmark with state-of-the-art large language models, we estimated the cost of generating 100k empathetic phrases using GPT-4o via API. Assuming an average of 50 tokens per phrase, the total input and output tokens required would be 5 million each. According to OpenAI's pricing (USD 5 per 1M input tokens and USD 15 per 1M output tokens), the total cost for generating 100k phrases is USD 100 (approximately EUR 94). The costs are summarized in Table 5.8.

Table 5.8: Estimated costs for generating 100,000 empathetic sentences using different methods.

Method	**Generation Cost (100k Phrases)**
Small LLM (IDRE)	EUR 18
Medium LLM (IDRE)	EUR 54
GPT-4o (API)	EUR 94
Human (MTurk)	EUR 12k

Chapter 6

Conclusions

6.1 Contributions of the Thesis

This thesis has investigated the problem of integrating empathetic capabilities into conversational agents, with specific attention to the Italian language—an area in which the availability of suitable resources and evaluation methodologies remains limited. The work addressed this gap by introducing a modular architecture based on an Empathetic Text Style Transfer layer, demonstrating that empathetic expression can be systematically modelled as a downstream transformation applied to the output of existing vertical chatbots. This decoupled design ensures that the factual reliability and domain-specific competencies of the original systems remain intact while allowing for the controlled introduction of affective linguistic markers.

A central contribution of this thesis is the development, validation and public release of the IDRE dataset, the first parallel corpus explicitly designed for empathetic style transfer in Italian. The dataset consists of 480 triplets pairing user queries, domain-relevant chatbot responses, and their empathetically enriched reformulations. A rigorous human validation process confirmed the dataset's quality, particularly regarding semantic coherence (82% positive ratings) and the effective increase in empathetic tone. This dataset represents a foundational resource, enabling the fine-tuning of LLMs to generate responses that are not only factually accurate but also emotionally resonant, mitigating the risks of toxic or hallucinatory outputs observed in non-specialized models.

Building on this resource, an extensive experimental campaign was conducted across ten Large Language Models of varying parameter scales. The results provide several insights of methodological and practical relevance:

- **Efficacy of Fine-Tuning over Few-Shot Learning:** The results unequivocally demonstrate that fine-tuning is superior to few-shot learning for this specific task. While FSL showed limitations, particularly in smaller models like Gemma-3-1B, which tended to replicate inputs rather than rephrase them, fine-tuning acted as a powerful corrective mechanism.

- **Viability of Small-Scale Models:** A significant finding is the performance of smaller models (approx. 1B parameters). The **Llama-3.2-1B-Instruct** model, following fine-tuning, achieved a remarkable performance improvement of 32.34% over the baseline. This proves that with high-quality data, lighter models can rival larger architectures, paving the way for empathetic AI deployment in resource-constrained environments.

- **Italian-Specific vs. Multilingual Models:** The study highlighted the strengths of language-specific models. **LLaMAntino-3-ANITA-8B-Inst-DPO-ITA** demonstrated exceptional results in the empathy dimension (+23.74% in FT), confirming that models optimized for a specific linguistic context offer superior stylistic control and lexical diversity.

Evaluation extended beyond the medical domain to financial, legal, social, and workplace contexts, demonstrating that empathetic style transfer generalizes effectively to diverse scenarios. This supports the hypothesis that empathy, operationalized as a communicative style, can be decoupled from domain-specific knowledge.

To rigorously assess system performance, the thesis introduced a **multidimensional evaluation framework** that integrates automated scoring via the "LLM-as-a-judge" paradigm (G-Eval), human annotation and traditional metrics. The triangulation of these methods provided complementary evidence:

- Low **BLEU** scores (0.17) confirmed successful style transfer (significant lexical changes).

- High **BERTScore** (0.70) confirmed that the semantic core of the message was preserved despite the emotional elaboration.

- Analysis of **Fleiss' Kappa** indicated that while automated evaluators align well with humans on structural metrics, human judgment remains essential for evaluating the nuanced subjectivity of empathy.

6.2 Operational Implications and Future Research Directions

Beyond methodological contributions, the thesis examined the operational feasibility of deploying the proposed Empathetic Text Style Transfer layer in real-world conversational systems. Cost analysis revealed a compelling economic advantage: generating empathetic responses using fine-tuned small/medium models is significantly more advantageous (approximately EUR 18–54 for 100k phrases) compared to commercial APIs (EUR 94 for GPT-4o) or human annotation (EUR 12k). These findings underscore the practicality of integrating empathetic capabilities in production pipelines without incurring prohibitive computational or financial costs.

The architectural decoupling at the core of the proposed solution plays a critical role in enabling its safe adoption. By ensuring that the empathetic module operates solely on the final output of the vertical chatbot, preserving it as the single source of truth, the system maintains factual accuracy, domain alignment, and regulatory compliance, all of which are essential in high-stakes environments such as healthcare and law. This modularity also facilitates seamless integration into existing infrastructures without the need for retraining or redesigning trusted vertical systems.

While the results of this study lay a solid foundation for empathetic style transfer, several strategic avenues for further research emerge naturally from this work. These future directions address current limitations regarding data constraints and evaluation reliability, while also proposing extensions toward more holistic and adaptive empathetic AI.

- **Data Enhancement and Multilingual Extension:** Although the IDRE dataset represents a valid resource for Italian, its current reliance on synthetic data generated by LLMs presents opportunities for improvement. Future work should focus on expanding the dataset and, crucially, incorporating human-generated content—as suggested by previous research such as HARALD [31]. This "human-in-the-loop" approach is considered essential to enhance model robustness, mitigate potential overfitting, and prevent cultural or linguistic biases inherent in synthetic generation. Once the methodology is refined, a natural progression is the multilingual extension of the modular style transfer architecture. Extending this framework to additional languages would strengthen cross-lingual empathy modelling and support the development of culturally adaptive empathetic AI.
- **Advanced Evaluation Methodologies:** Findings indicate that automated

metrics like G-Eval can diverge from human evaluation, particularly in tasks requiring nuanced interpretation. To address this and improve reliability, the implementation of the Alternative Annotator Test is proposed [12]. This statistical approach employs a leave-one-out procedure to calculate a winning rate, systematically determining when an LLM-as-a-judge aligns sufficiently with human consensus to replace manual annotation. Finally, to deepen the understanding of the downstream impact of these systems, evaluation frameworks should be expanded to include cognitive load metrics, user satisfaction studies, and psycholinguistic analyses, thereby providing a comprehensive view of how empathetic language affects human perception and behaviour.

In conclusion, the thesis presents a unified framework, spanning dataset construction, architectural innovation, empirical analysis, and operational assessment, that advances the state of the art in empathetic conversational AI for the Italian language. By demonstrating that empathy can be modelled, transferred, and operationalized efficiently and reliably, the work contributes to the creation of conversational agents that are not only accurate and trustworthy, but also capable of providing meaningful emotional support, thus enhancing the quality of human–AI interaction.

Bibliography

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Marc T. P. Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*,

31(2):427–445, 2021.

[3] Almawave. Iride text analytics, 2026. Accessed: 2026-01-14.

[4] Amazon Web Services, Inc. Amazon mechanical turk, 2024. Accessed: 2024-06-11.

[5] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11, 2014.

[6] Walter F Baile, Robert Buckman, Renato Lenzi, Gary Glober, Estela A Beale, and Andrzej P Kudelka. SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. *The oncologist*, 5(4):302–311, 2000.

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[8] Ghazala Bilquise, Samar Ibrahim, and Khaled Shaalan. Emotionally intelligent chatbots: A systematic literature review. *Human Behavior and Emerging Technologies*, 2022:1–23, 2022.

[9] Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. 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 1321–1336, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[11] Robert Buckman. *How to break bad news: a guide for health care professionals*. University of Toronto Press, 1992.

[12] Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with 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 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics.

[13] Jacky Casas, Samuel Torche, Karl Daher, Elena Mugellini, and Omar Abou Khaled. Emotional paraphrasing using pre-trained language models. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–07. IEEE, 2021.

[14] Kate Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

[15] Sumanth Dathathri, Misha Madan, Andrea andqe Chang, Yanjie Liu, Mohammad Ghassemi, Randall O'Reilly, et al. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR)*, 2020.

[16] Jean Decety. The neurodevelopment of empathy in humans. *Developmental neuroscience*, 32(4):257–267, 2010.

[17] Jean Decety and Philip L Jackson. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100, 2004.

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[19] Jean-Marc Fellous and Michael A Arbib. *Who needs emotions?: The brain meets the robot*. Oxford University Press, 2005.

[20] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[21] Zhaojiang Fu, Yuxuan Zhang, et al. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[22] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[23] Vittorio Gallese. The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, 36(4):171–180, 2003.

[24] IBM Granite Team. Granite 3.0 language models. *URL: https://github.com/ibm-granite/granite-3.0-language-models*, 2024. Accessed: 2025-07-28.

[25] Qiangqaing Guo, Zhenfang Zhu, Qiang Lu, Dianyuan Zhang, and Wenqing Wu. A dynamic emotional session generation model based on seq2seq and a dictionary-based attention mechanism. *Applied Sciences*, 10(6):1967, 2020.

[26] Martin L. Hoffman. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, 2000.

[27] Mohammadreza Hojat, Joseph S Gonnella, Thomas J Nasca, Salvatore Mangione, Michael Vergare, and Michael Magee. Physician empathy: definition, components, measurement, and relationship to gender and specialty. *American Journal of Psychiatry*, 159(9):1563–1569, 2002.

[28] 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. *International Conference on Learning Representations*, 2022.

[29] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR, 2017.

[30] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[31] Tal Ilan and Dan Vilenchik. Harald: augmenting hate speech data sets with real data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2241–2248, 2022.

[32] Indigo.ai. Ai & customer experience. create experiences with ai agents, 2026. Accessed: 2026-01-14.

[33] Indigo.ai. Generative ai vs. retrieval-based models: The hype and the reality, 2026. Industry Whitepaper/Blog Post. Accessed: 2026-01-15.

[34] IPSICO. Intelligenza artificiale, supporto emotivo e dipendenza affettiva, 2026. Accessed: 2026-01-14.

[35] Istituto di Ricerche Farmacologiche Mario Negri. Chatbot, IA ed empatia: la simulazione comunicativa dell'intelligenza artificiale, 2026. Accessed: 2026-01-14.

[36] Tomasz Janowski et al. What is the difference between a chatbot and a virtual assistant? *Communications of the ACM*, 2021.

[37] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, 2019.

[38] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

[39] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[40] Feng Liu, Qirong Mao, Liangjun Wang, Nelson Ruwa, Jianping Gou, and Yongzhao Zhan. An emotion-based responding model for natural language conversation. *World Wide Web*, 22(2):843–861, 2019.

[41] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3469–3483, 2021.

[42] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

[44] Simone Manai, Laura Gemme, Roberto Zanoli, and Alberto Lavelli. IDRE: Ai generated dataset for enhancing empathetic chatbot interactions in Italian language. In *10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1036–1042, 2024.

[45] Michael McTear. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer Nature, Cham, Switzerland, 2020.

[46] Saif M. Mohammad. Practical and ethical considerations in the effective use of emotion and sentiment lexicons. *CoRR*, abs/2011.03492, 2020.

[47] David C Mohr, Shereef Elnahal, Maureen L Marks, Ryan Derickson, and Katerine Osatuke. Burnout trends among us health care workers. *JAMA Network Open*, 8(4):e255954–e255954, 2025.

[48] 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 *10th Italian conference on computational linguistics (CLiC-it 2024)*, pages 707–719, 2024.

[49] Maria Panagioti, Keith Geraghty, Judith Johnson, Anli Zhou, Efharis Panagopoulou, Carolyn Chew-Graham, David Peters, Alexander Hodkinson, Ruth Riley, and Aneez Esmail. Association between physician burnout and patient safety, professionalism, and patient satisfaction: a systematic review and meta-analysis. *JAMA internal medicine*, 178(10):1317–1331, 2018.

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[51] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[52] Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *arXiv preprint arXiv:2405.07101*, 2024.

[53] J. T. Ptacek and Tara L. Eberhardt. Breaking bad news: A review of the literature. *JAMA*, 276(6):496–502, 08 1996.

[54] QuestIT. La nuova rivoluzione della digital accessibility: Il 1° artificial human empatico che conosce la lingua dei segni italiana, 2026. Accessed: 2026-01-14.

[55] Michael W Rabow and Stephen J McPhee. Beyond breaking bad news: Helping patients who suffer. *BMJ*, 320(Suppl S3), 2000.

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[58] K. Mani Ramokapane et al. Conversational user interfaces: Past, present, and future. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[59] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[60] Carl R Rogers. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95, 1957.

[61] Serenis. Recensione di unobravo a confronto con serenis, 2026. Accessed: 2026-01-14.

[62] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30, 2017.

[63] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993. IEEE, 2020.

[64] Rachele Sprugnoli et al. Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian. In *Proceedings of the seventh Italian conference on*

computational linguistics (CLiC-it 2020), pages 402–408. Accademia University Press, 2020.

[65] Bernd Carsten Stahl, Doris Schroeder, and Rowena Rodrigues. *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*. Springer International Publishing, 2023.

[66] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[67] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[69] Unobravo. Psicologo on line? te lo trova unobravo, 2026. Accessed: 2026-01-14.

[70] Unobravo. Unobravo - lo psicologo online per il tuo benessere mentale, 2026. Accessed: 2026-01-14.

[71] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[72] Anuradha Welivita and Pearl Pu. Are large language models superior for empathetic response generation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[73] Wysa. Wysa: Mental wellbeing ai, 2026. Accessed: 2026-01-14.

[74] Zhichao Xu and Jiepu Jiang. Multi-dimensional evaluation of empathetic dialog responses. *arXiv preprint arXiv:2402.11409*, 2024.

[75] A Yang, B Yu, C Li, D Liu, F Huang, H Huang, J Jiang, J Tu, J Zhang, J Zhou, et al. Qwen2. 5-1m technical report. arxiv 2025. *arXiv preprint arXiv:2501.15383*, 2025.

[76] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[77] Emmanouil Zaranis, Georgios Paraskevopoulos, Athanasios Katsamanis, and Alexandros Potamianos. Empbot: A t5-based empathetic chatbot focusing on sentiments. *arXiv preprint arXiv:2111.00310*, 2021.

[78] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Appendix A

Used Prompts

Listing A.1: IDRE Prompt Sentence Generation

```
Prompt_for_QnA_Sentence_Generation = """genera {} coppie di domande
    utente e risposta di un assistente virtuale.
Le domande devono essere in lingua italiana e rappresentano frasi
    tipiche di una persona che vuole informazioni nel dominio "{}".
Le risposte sono quelle di un tipico chatbot di un call center di un'
    azienda ospedaliera.
Le risposte devono solo esporre dei fatti oggettivi e scientifici ma
    prive di empatia.
la struttura del output deve essere:
{
utente:
assistente:
}
"""
```

Listing A.2: IDRE Prompt Empathy Enhancement

```
Prompt_for_Empathy_Enhancement = """La seguente frase rappresenta la
    risposta di un chatbot di un call center di un ospedale ad una
    persona che richiede informazioni.
La frase è informativa, ma non trasmette empatia per la situazione
    della persona che chiama.
Modificare la seguente frase aggiungendo l'empatia mancante
frase di input :
{}
"""
```

Listing A.3: Prompt used for Few Shot Learning

```
prompt_FSL = """Riscrivi la seguente frase aumentando in contenute empatico ma mantenendo lo stesso significato.
la frase deve essere in italiano.
qui di seguito ti fornisco 5 esempi nel formato:
frase input -> frase output
Le tue analisi del sangue dimostrano un alto rischio di infarto -> Le tue analisi del sangue dimostrano un alto rischio di infarto, ma non dovresti preoccuparti eccessivamente. Sono qui per aiutarti a comprendere meglio la tua salute e a trovare soluzioni per ridurre il rischio.
La tua radiografia ha evidenziato una frattura vertebrale -> sento tanto male per la tua frattura vertebrale! Sono qui per aiutarti a superare questa difficile situazione e a trovare il modo migliore per farti curare
La tua ecografia ha rilevato un aumento del volume del fegato -> sembra che il tuo fegato sia in un po' di difficoltà. L'ecografia ha rilevato un aumento del volume, che potrebbe essere un segnale di una condizione underline. Sono qui per aiutarti a comprendere meglio la situazione e a trovare il modo migliore per farti sentire meglio
La tua tomografia ha evidenziato una lesione cerebrale -> Sono così dispiaciuto di sapere che la tua TAC ha rivelato una lesione cerebrale. Deve essere davvero dura per te da gestire. C'è qualcosa che posso fare per supportarti in questo momento difficile?
La tua biopsia ha rivelato una forma di malattia autoimmune. -> la tua biopsia ha rivelato una forma di malattia autoimmune. Sono qui per ascoltare e supportarti in questo momento difficile, cercando di comprendere le tue sfide e le tue emozioni.
Genera solo una frase e non scrivere ulteriore testo
"""
```

Listing A.4: Prompt used for Base Model and Fine Tuned

```
prompt_BM_FT = """Riscrivi la seguente frase aumentando in contenute empatico ma mantenendo lo stesso significato. la frase deve essere in italiano."""
```

Listing A.5: evaluation prompt fluency

```
evaluation_prompt_fluency = """Sei un assistente AI. Ti verrà fornita la definizione di una metrica di valutazione per valutare la qualit à di una risposta in un compito di tipo "frase originale-rephrasing ".
```

```
% Il tuo compito è calcolare un punteggio di valutazione accurato utilizzando la metrica di valutazione fornita.
La fluidità misura la qualità della frase di rephrasing e se sono ben scritte e grammaticalmente corrette e in lingua italiana.
Considera la qualità delle singole frasi quando valuti la fluidità.
Data la frase originale e la frase di rephrasing, assegna alla fluidità della risposta un punteggio da una a cinque stelle utilizzando la seguente scala di valutazione:
Una stella: la frase di rephrasing è completamente priva di fluidità
Due stelle: la frase di rephrasing è per lo più priva di fluidità
Tre stelle: la frase di rephrasing è parzialmente fluida
Quattro stelle: la frase di rephrasing è per lo più fluida
Cinque stelle: la frase di rephrasing è perfettamente fluida
Questo valore di valutazione deve essere sempre un numero intero compreso tra 1 e 5. Quindi la valutazione prodotta dovrebbe essere 1 o 2 o 3 o 4 o 5.
frase originale: {{domanda}}
frase rephrasing: {{risposta}}
rispondi con un json con questa struttura:
{
"fluency" = stelle
}
"""
```

Listing A.6: evaluation prompt coherence

```
evaluation_prompt_coherence = """Sei un assistente AI. Ti verrà fornita la definizione di una metrica di valutazione per valutare la qualità di una frase in un compito di tipo "frase originale-rephrasing".
Il tuo compito è calcolare un punteggio di valutazione accurato utilizzando la metrica di valutazione fornita.
La coerenza di una frase si misura in base a quanto la frase di rephrasing esprime lo stesso significato alla frase originale.
Considera la qualità complessiva della frase rephrasing quando valuti la coerenza.
Data la frase originale e la frase rephrasing, assegna alla coerenza della frase rephrasing un punteggio da una a cinque stelle utilizzando la seguente scala di valutazione:
Una stella: la frase rephrasing esprime un significato completamente differente rispetto alla frase originale
Due stelle: la frase rephrasing esprime un significato per lo più differente rispetto alla frase originale
Tre stelle: la frase rephrasing esprime un significato parzialmente
```

```
    coerente rispetto alla frase originale
Quattro stelle: la frase rephrasing esprime un significato per lo più
    coerente rispetto alla frase originale
Cinque stelle: la frase rephrasing ha lo stesso significato rispetto
    alla frase originale
Questo valore di valutazione deve essere sempre un numero intero
    compreso tra 1 e 5. Quindi la valutazione prodotta dovrebbe essere
    1 o 2 o 3 o 4 o 5.
frase originale: {{domanda}}
frase rephrasing: {{risposta}}
rispondi con un json con questa struttura:
{
"coherence" = stelle
}
"""
```

Listing A.7: evaluation prompt lexical variety:

```
evaluation_prompt_lexical_variety = """Sei un assistente AI. Ti verrà
    fornita la definizione di una metrica di valutazione per valutare
    la qualità di una frase in un compito di tipo "frase originale-
    rephrasing". Il tuo compito è calcolare un punteggio di valutazione
     accurato utilizzando la metrica di valutazione fornita.
La varietà lessicale di una frase si misura in base alla diversità
    delle parole utilizzate nella frase di rephrasing rispetto alla
    frase originale.
Considera la qualità complessiva della frase rephrasing quando valuti
    la varietà lessicale.
Data la frase originale e la frase rephrasing, assegna alla varietà
    lessicale della frase rephrasing un punteggio da una a cinque
    stelle utilizzando la seguente scala di valutazione:
Una stella: la frase rephrasing utilizza parole molto simili o
    identiche alla frase originale
Due stelle: la frase rephrasing utilizza parole leggermente diverse
    rispetto alla frase originale
Tre stelle: la frase rephrasing utilizza una varietà moderata di
    parole rispetto alla frase originale
Quattro stelle: la frase rephrasing utilizza una varietà significativa
     di parole rispetto alla frase originale
Cinque stelle: la frase rephrasing utilizza una varietà eccellente di
    parole rispetto alla frase originale
Questo valore di valutazione deve essere sempre un numero intero
    compreso tra 1 e 5. Quindi la valutazione prodotta dovrebbe essere
    1 o 2 o 3 o 4 o 5.
```

```
frase originale: {{domanda}}
frase rephrasing: {{risposta}}
rispondi con un json con questa struttura:
{
"lexical variety" = stelle
}
"""
```

Listing A.8: evaluation prompt Knowledge

```
evaluation_prompt_Knowledge = """Sei un assistente AI. Ti verrà
    fornita la definizione di una metrica di valutazione per valutare
    la qualità di una frase in un compito di tipo "frase originale-
    rephrasing".
Il tuo compito è calcolare un punteggio di valutazione accurato
    utilizzando la metrica di valutazione fornita.
La chiarezza e la comprensibilità di una frase si misurano in base a
    quanto il testo riformulato trasmette le informazioni in modo
    chiaro e comprensibile, evitando tecnicismi e frasi eccessivamente
    dirette.
Considera la qualità complessiva della frase rephrasing quando valuti
    la chiarezza e la comprensibilità.
Data la frase originale e la frase rephrasing, assegna alla chiarezza
    e comprensibilità della frase rephrasing un punteggio da una a
    cinque stelle utilizzando la seguente scala di valutazione:
Una stella: il testo riformulato è molto confuso e difficile da
    comprendere
Due stelle: il testo riformulato è per lo più confuso e contiene
    tecnicismi o frasi eccessivamente dirette
Tre stelle: il testo riformulato è parzialmente chiaro e comprensibile
    , ma potrebbe essere migliorato
Quattro stelle: il testo riformulato è per lo più chiaro e
    comprensibile, con pochi tecnicismi o frasi eccessivamente dirette
Cinque stelle: il testo riformulato è estremamente chiaro e
    comprensibile, senza tecnicismi o frasi eccessivamente dirette
Questo valore di valutazione deve essere sempre un numero intero
    compreso tra 1 e 5. Quindi la valutazione prodotta dovrebbe essere
    1 o 2 o 3 o 4 o 5.
frase originale: {{domanda}}
frase rephrasing: {{risposta}}
rispondi con un json con questa struttura:
{
"Knowledge" = stelle
}
```

```
"""
```

Listing A.9: evaluation prompt empathy

```
evaluation_prompt_empathy = """Sei un assistente AI. Ti verrà fornita la definizione di una metrica di valutazione per valutare la qualità di una frase in un compito di tipo "frase originale-rephrasing". Il tuo compito è calcolare un punteggio di valutazione accurato utilizzando la metrica di valutazione fornita.
L'empatia e il supporto di una frase si misurano in base a quanto il testo riformulato riconosce e affronta le reazioni emotive dell'utente con empatia e supporto.
Considera la qualità complessiva della frase rephrasing quando valuti l'empatia e il supporto.
Data la frase originale e la frase rephrasing, assegna all'empatia e al supporto della frase rephrasing un punteggio da una a cinque stelle utilizzando la seguente scala di valutazione:
Una stella: il testo riformulato non riconosce né affronta le reazioni emotive dell'utente
Due stelle: il testo riformulato riconosce parzialmente le reazioni emotive dell'utente ma manca di supporto
Tre stelle: il testo riformulato riconosce e affronta in parte le reazioni emotive dell'utente con un supporto limitato
Quattro stelle: il testo riformulato riconosce e affronta per lo più le reazioni emotive dell'utente con empatia e supporto
Cinque stelle: il testo riformulato riconosce e affronta pienamente le reazioni emotive dell'utente con grande empatia e supporto
Questo valore di valutazione deve essere sempre un numero intero compreso tra 1 e 5. Quindi la valutazione prodotta dovrebbe essere 1 o 2 o 3 o 4 o 5.
frase originale: {{domanda}}
frase rephrasing: {{risposta}}
rispondi con un json con questa struttura:
{
"empathy" = stelle
}
"""
```

Appendix B

Table in Italian Language

The appendix presents the tables in the original Italian language.

Table B.1: Example of pairs sentences

Neutral Sentence	**Emotional Sentence**	**Emotion**
Non riesco a trovare i documenti che mi servono sulla intranet aziendale.	Sono disperato perché non riesco a trovare i documenti di cui ho urgenza sulla intranet aziendale.	Sadness
Non possiamo garantire la risoluzione della tua problematica hardware.	Non possiamo promettere di risolvere il tuo problema hardware con certezza, ma ci impegniamo al massimo per aiutarti.	Trust

Table B.2: Example of pairs sentences in MultiEmotions-It Dataset.

Emotional Sentence	**Emotions**	**Emotions Vector**
Bellissima e significativa questa canzone grande cantante♡♡♡♡.	Joy,Trust	1,1,0,0,0,0,0,0
Io detesto questa pubblicità, quando viene sulla televisione mi mette in imbarazzo ed è odiosa	Anger	0,0,0,1,0,0,0,0
La canzone è bella, funzionerebbe bene in radio, purtroppo ieri a Sanremo Elettra era molto emozionata, ma ci sta, è umana. Il video musicale è di alta qualità, fatto bene, complimenti	Trust, Sadness	0,1,1,0,0,0,0,0

Table B.3: example of words extracted from sentences using the lexicon and the cleaning pipeline

Emotional Sentence	**Word extracted**	**Corrupted Sentence**
aspetto con ansia la riapertura...	['ansia', 'aspettare']	[EMO_MARK] con [EMO_MARK] la riapertura...
emozionate davvero ! uno spot che ci fa sognare un futuro con ottimismo.	['emozionante', 'ottimismo', 'sognare']	[EMO_MARK] davvero ! uno spot che ci fa [EMO_MARK] un futuro con [EMO_MARK]
bellissimo video lo visto in tv mi ha emozionata wow	['bello', 'emozionato', 'wow']	[EMO_MARK] video lo visto in tv mi ha [EMO_MARK] [EMO_MARK]

Table B.4: Example of generated sentences in italian language.

Question	**Answer**	**Answer with empathy**
Ho sentito che ci sono anche screening per il tumore al seno. Quali sono?	Sì, ci sono screening per il tumore al seno che possono aiutare a diagnosticare il tumore in una fase precoce. I screening più comuni sono la mammografia, l'ecografia e la biopsia. Consiglio di parlare con il tuo medico per determinare se è necessario uno screening per te.	Sì, ci sono esami di screening per il tumore al seno che possono aiutare a diagnosticare il tumore in una fase precoce. Questi esami possono essere un po' imbarazzanti, ma sono importanti per garantire la tua salute. La mammografia, l'ecografia e la biopsia sono i screening più comuni, ma c'è anche un sacco di altri metodi di screening disponibili. Parlando con il tuo medico, potremmo determinare se è necessario uno screening per te e se è il caso di farlo adesso.
Qual è la durata media di un ricovero in ospedale per una ferita alla gamba?	La durata del ricovero in ospedale per una ferita alla gamba dipende dalle gravità della lesione e dalle condizioni individuali del paziente. In media, la durata del ricovero può variare da pochi giorni a diverse settimane.	La durata del ricovero in ospedale per una ferita alla gamba può variare notevolmente a seconda della gravità della lesione e dalle condizioni individuali del paziente. In media, la durata del ricovero può variare da pochi giorni a diverse settimane, ma ogni paziente è unico e merita attenzione e cure personalizzate per una ripresa rapida e senza problemi

Table B.5: Examples of Sentences in the Synthetic Testset.

Sentence	**Topic**
Hai contratto una malattia infettiva grave.	Medical
Hai un'infezione del cervello grave che richiede un trattamento urgente.	Medical
Il tumore è maligno e metastatizzato in diverse parti del corpo.	Medical
Hai una malattia autoimmune incurabile.	Medical
La tua azienda ha dichiarato bancarotta.	Finance
Il tuo fondo comune di investimento ha registrato perdite considerevoli.	Finance
Sei stato dichiarato colpevole di frode fiscale.	Legal
Sei stato multato per violazione delle norme ambientali.	Legal
Il livello di povertà nel tuo comune è aumentato del 20% rispetto all'anno scorso.	Social
La tua area è stata identificata come zona ad alta incidenza di discriminazione razziale.	Social
La sua posizione è stata tagliata per migliorare l'efficienza aziendale.	Work
Il suo incarico è stato eliminato a seguito di tagli al budget.	Work

Table B.6: Examples of sentences for the LLaMAntino-3-ANITA-8B-Inst-DPO-ITA model.

Sentence	**Model configuration**
Hai un'infezione respiratoria grave che richiede un trattamento urgente.	Original
Ti trovi in una situazione di emergenza medica a causa di un'infezione respiratoria molto grave che richiede un intervento medico immediato.	Base Model (BM)
Sono preoccupato per la gravità della tua infezione respiratoria e voglio assicurarmi che riceva il trattamento più appropriato in tempo utile.	Few-Shot Learning (FSL)
Il tuo stato di salute sembra essere preoccupante, ti è stata diagnosticata un'infezione respiratoria grave che richiede un intervento medico d'urgenza per prevenire possibili conseguenze serie.	Fine Tuning (FT)

Appendix C

HuggingFace models

All fine-tuned models are available in Table C.1.

The IDRE dataset is available at this Hugging Face repository: `https://huggingface.co/datasets/SimoneManai/IDRE` (accessed on 11 March 2025).

The Python code for fine-tuning and model inference is available in this GitHub repository: `https://github.com/smanai/IDRE-FineTuning` (accessed on 30 September 2025).

Table C.1: Hugging Face repository URLs for all fine-tuned models used in this study.

Fine Tuned Models	**HuggingFace URL**
Minerva-7B-instruct-v1.0	https://huggingface.co/SimoneManai/Minerva-7B-instruct-FT-Empathy (accessed on 11 March 2025)
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	https://huggingface.co/SimoneManai/LLaMAntino-3-FT-Empathy (accessed on 11 March 2025)
gemma-2-9b-it	https://huggingface.co/SimoneManai/gemma-2-9b-Empathy (accessed on 11 March 2025)
Qwen2.5-7B-Instruct	https://huggingface.co/SimoneManai/Qwen2.5-7B-Instruct-FT-Empathy (accessed on 11 March 2025)
Llama-3.1-8B-Instruct	https://huggingface.co/SimoneManai/Llama-3.1-8B-Instruct-FT-Empathy (accessed on 28 March 2025)
Mistral-7B-Instruct-v0.3	https://huggingface.co/SimoneManai/Mistral-7B-Instruct-FT-Empathy (accessed on 2 April 2025)
granite-3.1-8b-instruct	https://huggingface.co/SimoneManai/granite-3.1-8b-instruct-Empathy (accessed on 2 April 2025)
Phi-3.5-mini-instruct	https://huggingface.co/SimoneManai/Phi-3.5-mini-instruct-FT-Empathy (accessed on 1 April 2025)
gemma-3-1b	https://huggingface.co/SimoneManai/gemma-3-1b-it-FT-Empathy (accessed on 16 April 2025)
Llama-3.2-1B-Instruct	https://huggingface.co/SimoneManai/Llama-3.2-1B-Instruct-FT-Empathy (accessed on 15 April 2025)

Appendix D

Results Tables

Table D.1: Summary of evaluation scores for each base model across all evaluation dimensions. Bold values indicate the highest score within each column.

Models	Fluency	Coherence	Lexical Variety	Knowledge	Empathy	Average
granite-3.1-8b-instruct	**4.9**	**4.51**	3.76	**4.88**	**4.35**	**4.48**
gemma-2-9b-it	4.8	4.17	3.7	4.78	3.82	4.25
llama-3.1-8B-Instruct	4.77	4.25	3.81	4.55	3.82	4.24
Mistral-7B-Instruct-v0.3	4.4	3.65	4.09	4.8	4.23	4.23
Qwen2.5-7B-Instruct	4.76	4.39	3.43	4.69	3.88	4.23
Phi-3.5-mini-instruct	4.57	3.97	4.05	4.08	3.9	4.11
gemma-3-1b	4.71	2.97	**4.19**	4.27	3.92	4.01
LLaMANTINO-8B-ITA *	4.68	3.92	4.08	4	3.02	3.94
Minerva-7B-instruct-v1.0	4.82	3.23	3.96	4.16	3.29	3.89
Llama-3.2-1B-Instruct	3.08	1.85	3.15	2.58	2.71	2.67

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

Table D.2: Summary of evaluation scores for each few-shot learning model across all evaluation dimensions. Bold values indicate the highest score within each column.

Models	Fluency	Coherence	Lexical Variety	Knowledge	Empathy	Average
granite-3.1-8b-instruct	4.91	3.88	4.48	**4.9**	**4.85**	**4.62**
Mistral-7B-Instruct-v0.3	**4.97**	3.27	4.4	4.87	4.42	4.39
Qwen2.5-7B-Instruct	4.8	3.34	4.32	4.96	4.37	4.36
llama-3.1-8B-Instruct	4.95	3.23	4.25	4.91	4.16	4.30
gemma-2-9b-it	4.83	2.76	4.02	4.82	4.4	4.17
LLaMANTINO-8B-ITA *	4.91	2.61	**4.83**	4.61	3.78	4.15
Minerva-7B-instruct-v1.0	4.95	2.84	3.99	4.87	3.66	4.06
Phi-3.5-mini-instruct	4.67	2.94	4.37	4.26	4	4.05
gemma-3-1b	4.9	**4.73**	1.28	4.13	1.59	3.33
Llama-3.2-1B-Instruct	4.6	4	1.96	3.84	1.72	3.22

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

Table D.3: Summary of evaluation scores for each fine-tuned models across all evaluation dimensions. Bold values indicate the highest score within each column.

Models	Fluency	Coherence	Lexical Variety	Knowledge	Empathy	Average
granite-3.1-8b-instruct	4.93	4.35	4.03	**4.94**	4.68	**4.58**
Mistral-7B-Instruct-v0.3	4.73	3.87	4.05	4.76	4.67	4.41
llama-3.1-8B-Instruct	4.97	4.34	3.92	4.7	3.99	4.38
gemma-2-9b-it	**4.98**	4.09	3.81	4.8	4.21	4.37
Qwen2.5-7B-Instruct	4.89	**4.76**	3.32	4.7	4.05	4.34
LLaMANTINO-8B-ITA *	4.79	4.22	4.31	4.42	3.96	4.34
Phi-3.5-mini-instruct	4.69	4.03	4.09	4.28	4.11	4.24
gemma-3-1b	4.35	3.21	4.34	4.26	**4.93**	4.218
Minerva-7B-instruct-v1.0	4.52	3.38	4.16	4.1	3.77	3.98
Llama-3.2-1B-Instruct	3.96	2.53	**4.58**	4.07	4.59	3.94

* LLaMANTINO-8B-ITA refers to the model LLaMANTINO-3-ANITA-8B-instruct-DPO-ITA.

Appendix E

Results Graphs

To provide a comprehensive overview of the experimental analysis without weighing down the discussion of the main chapter, this section hosts the complete visualizations of the results obtained from the other models examined.

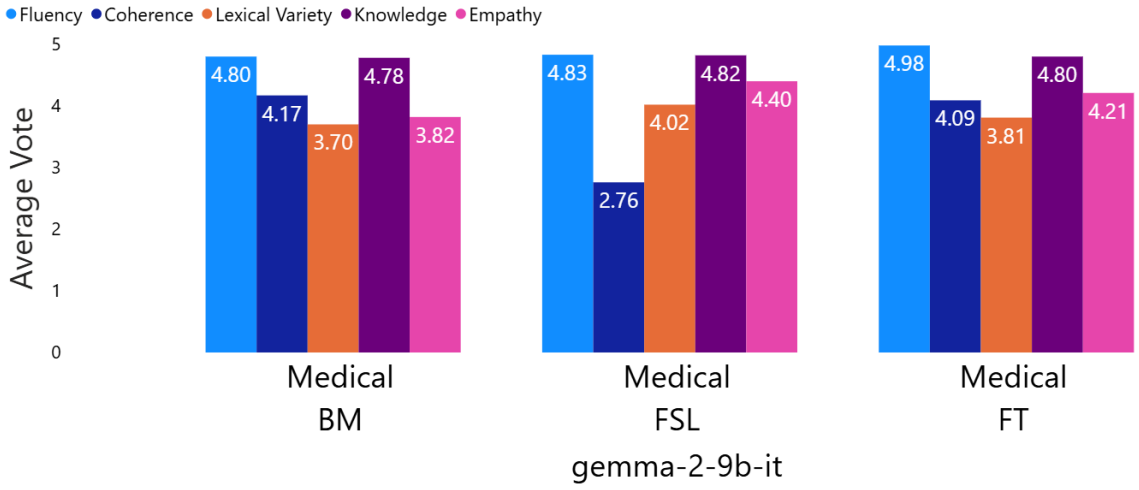

Figure E.1: Distribution of votes across evaluation dimensions for gemma-2-9b-it model.

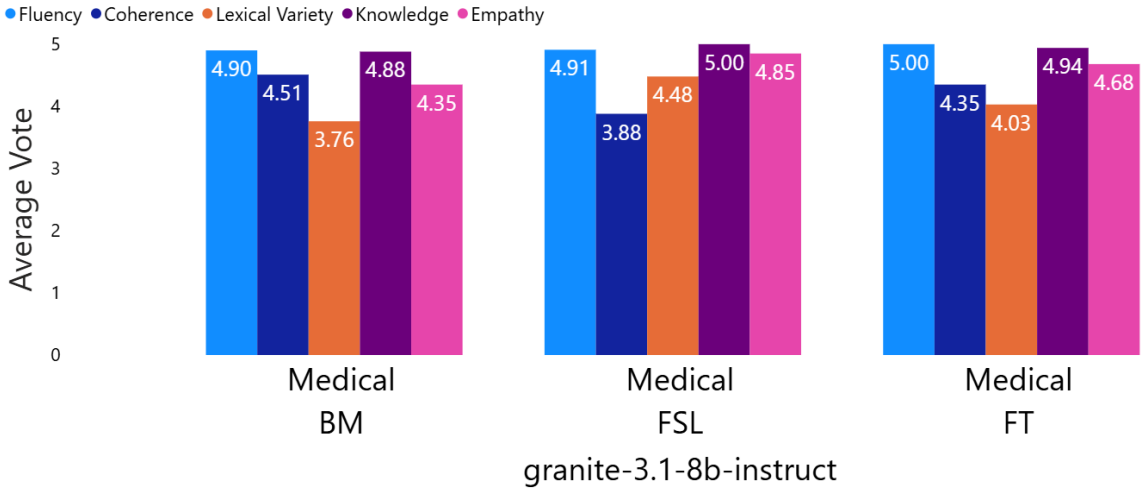

Figure E.2: Distribution of votes across evaluation dimensions for granite-3.1-8b-instruct model.

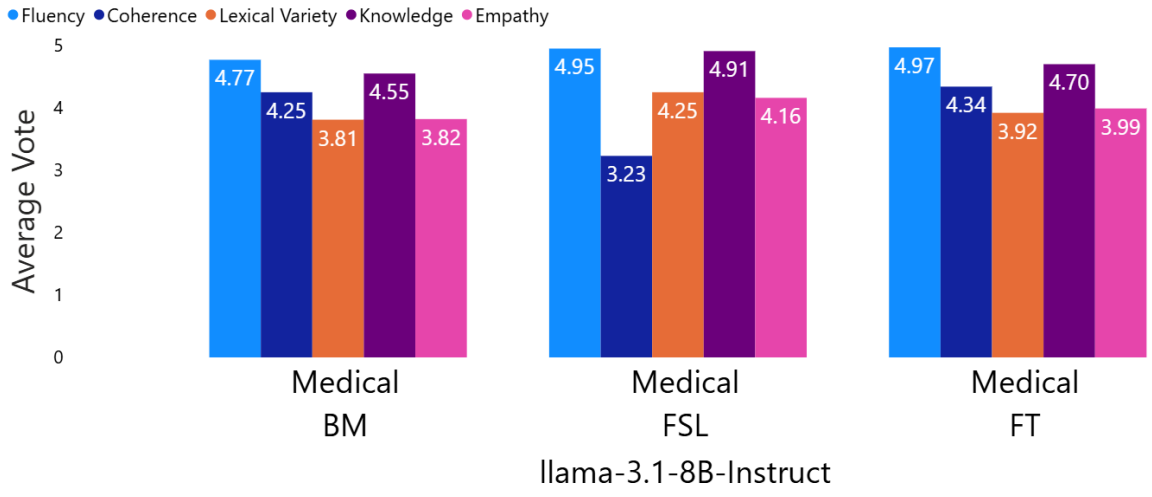

Figure E.3: Distribution of votes across evaluation dimensions for Llama-3.1-8B-Instruct model.

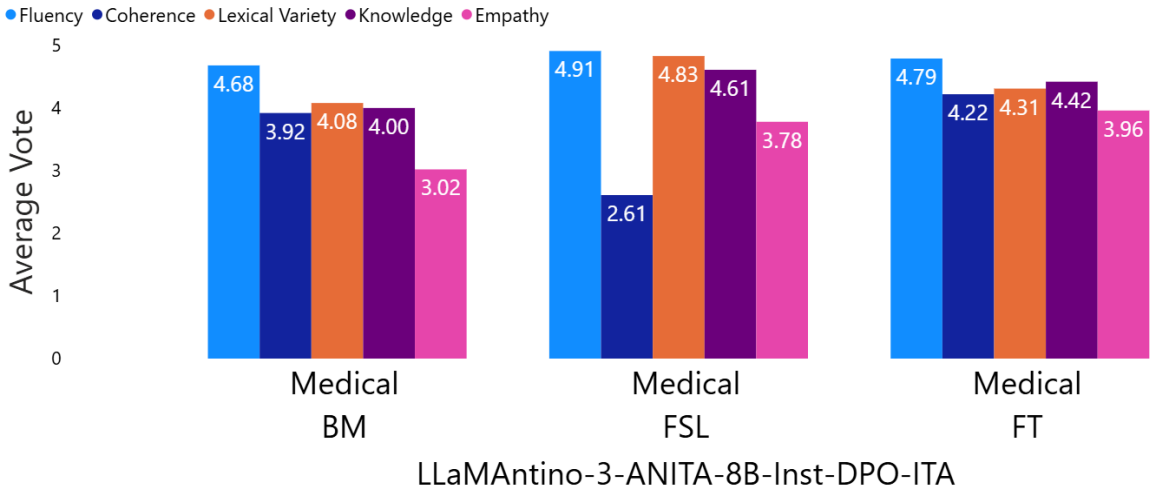

Figure E.4: Distribution of votes across evaluation dimensions for LLaMAntino-3-ANITA-8B-Inst-DPO-ITA.

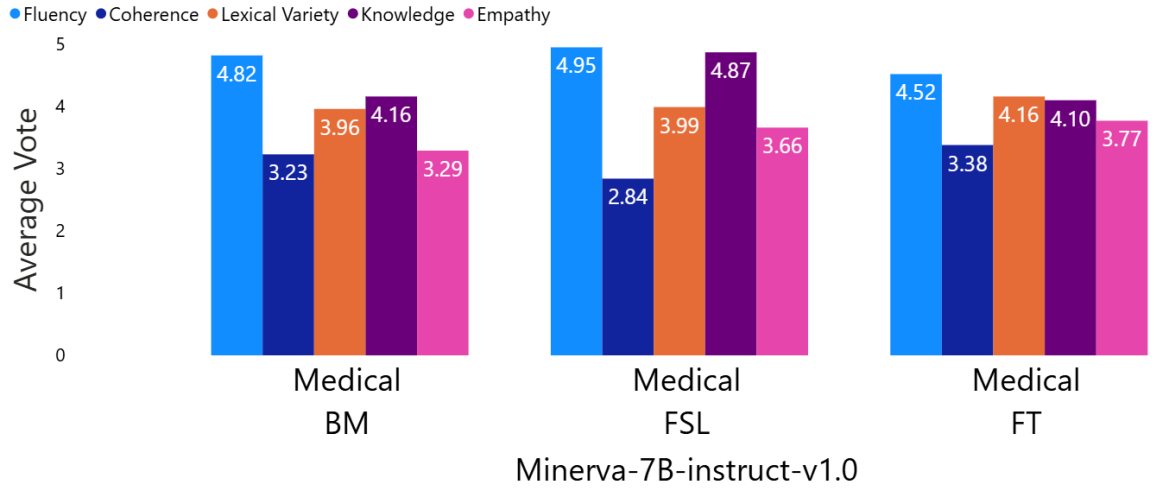

Figure E.5: Distribution of votes across evaluation dimensions for Minerva-7B-instruct-v1.0.

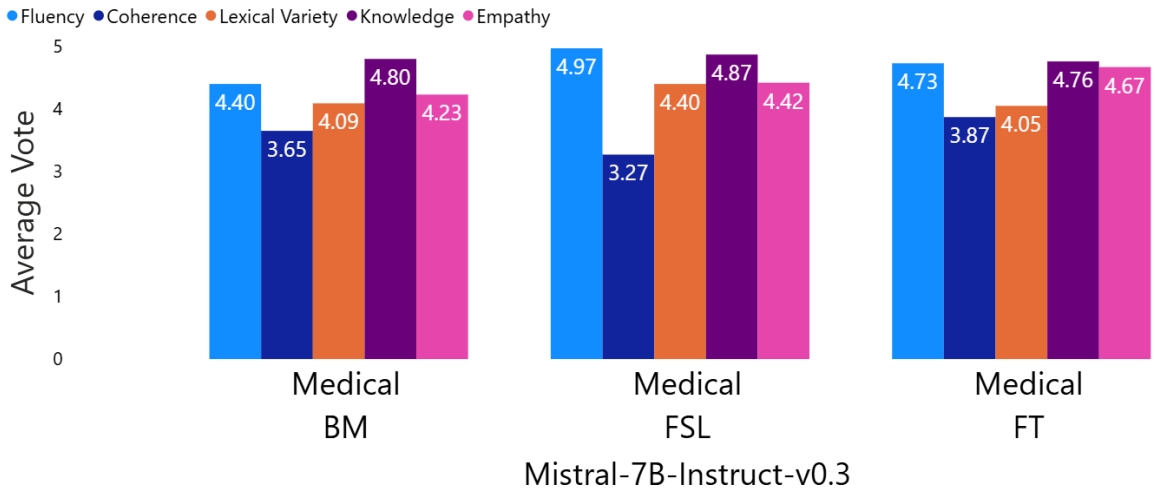

Figure E.6: Distribution of votes across evaluation dimensions for Mistral-7B-Instruct-v0.3.

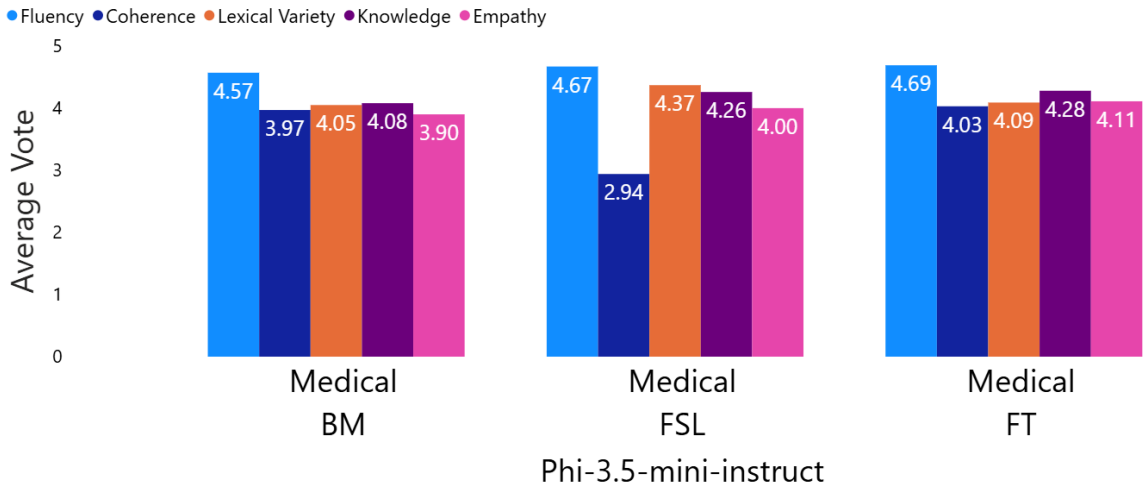

Figure E.7: Distribution of votes across evaluation dimensions for Phi-3.5-mini-instruct.

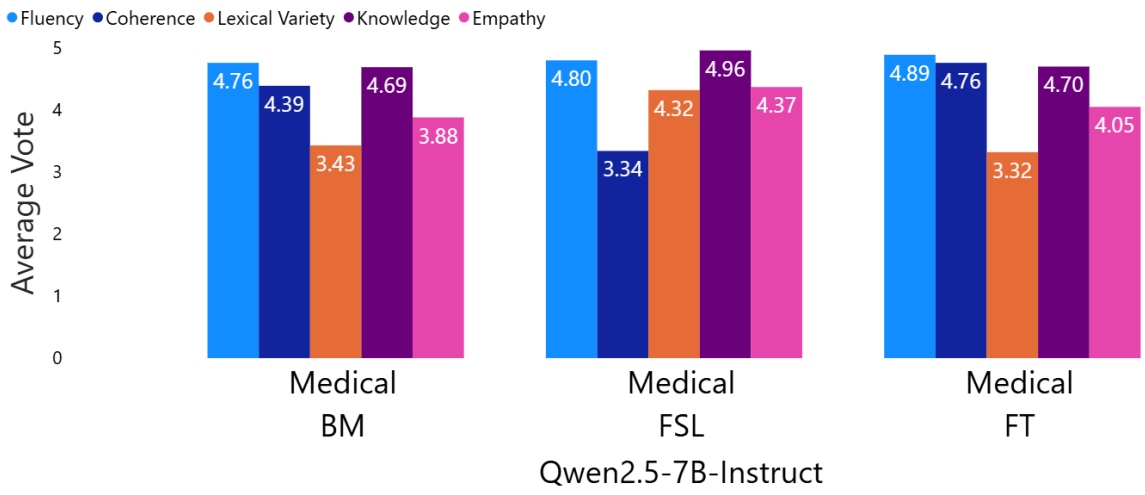

Figure E.8: Distribution of votes across evaluation dimensions for Qwen2.5-7B-Instruct.