UNIVERSITÀ
DI TRENTO

Department of Information Engineering and Computer Science

ICT International Doctoral School

~ · ~

38th Cycle

Through Smoke and Mirrors of the Post-Truth Era

Knowledge-Driven Generation as Algorithmic Resistance to Misinformation

Advisor
Dr. Marco Guerini

Co-Advisor
Prof. Jacopo Staiano

Candidate
Daniel Russo

July 10, 2026

Abstract

The rapid spread of misinformation online poses a growing threat to public discourse and democratic society. Research has shown that simply flagging false content is insufficient; providing users with well-grounded explanations of why a claim is false leads to more durable belief revision than mere labeling. Yet the volume at which misinformation spreads online far exceeds the capacity of manual fact-checking, making its automation not merely desirable but necessary. Natural Language Generation (NLG) represents a viable solution, enabling the automatic production of verdicts: knowledge-grounded textual responses that explain the veracity of a claim. This thesis investigates how knowledge-driven generative approaches can automate verdict generation, leveraging the growing capabilities of modern NLG technologies while accounting for the multifaceted nature of misinformation, including its diverse communication styles, linguistic and geographical spread, and the varying availability of reliable knowledge in real-world scenarios. We first cast verdict generation as a summarization task, benchmarking extractive, abstractive, and hybrid approaches and exploring multitask strategies to adapt verdict style to the guidelines of different fact-checking organizations. While effective, these systems were developed and evaluated on journalistic claims, failing to account for the social media context in which much misinformation circulates. We therefore shift to a social correction scenario, constructing a dedicated dataset and extending verdict generation to jointly adapt style and emotional register to the communication patterns of users spreading false claims. Both studies, however, remain limited to English and rely on knowledge available in the same language as the claim. We address this limitation by leveraging multilingual instruction-tuned large language models, building a professionally curated dataset across eight European languages and evaluating cross-lingual scenarios where supporting knowledge and claims are in different languages. We then turn to the more realistic setting in which no fact-checking article is directly associated with the claim, evaluating Retrieval-Augmented Generation pipelines that retrieve evidence from both homogeneous and heterogeneous knowledge bases, testing diverse retrieval strategies and examining how claim style affects pipeline performance. Finally, we recognize that not all misinformation takes the form of explicit, well-formed claims: clickbait headlines deceive through omission and require dedicated preprocessing before they can enter a fact-checking pipeline. We address this gap by proposing novel tasks and purpose-built resources for clickbait detection, spoiler generation, and neutralization. The thesis concludes by discussing the open challenges that remain in the context of counterspeech for misinformation in NLG.

Acknowledgements

Although the following pages might provide a comprehensive picture of my PhD years, there is no way to give a truly exhaustive account of this journey I decided to embark on, almost naively, years ago. What these pages will hide are the dozens of wrong paths I took, the dead ends, the right and wrong decisions that, together, made me realise that what is supposed to be a mere three-year research project is, in fact, a personal journey, an uphill battle. If this climb has come to an end, it is because of all the people who accompanied me, who shared all, or part, of this journey with me; acknowledging them in this brief text is the least I can do to express my gratitude.

First and foremost, if I can finally say out loud "This is real, this is it!" it is thanks to my advisor, Dr. Marco Guerini. There will never be enough words to express how grateful I am for his believing in me since the very beginning, and for granting me opportunities that allowed me to grow as a researcher and as a person. Thanks for guiding me through this journey, making sure I kept climbing, while always leaving me the freedom to explore, to make mistakes, and to retrace my steps. No less importantly, Prof. Jacopo Staiano, my co-advisor, had a profound impact on the success of this venture, and I cannot be more grateful for that. And finally, Prof. Oscar Araque, who welcomed me into the GSI group at the *Universidad Politécnica de Madrid*, and embraced my last-minute crazy ideas.

Another important piece of this journey is my group, who shared with me, daily, every joy and sorrow: thank you for being the best fellow travellers I could have ever asked for. I would also like to extend my gratitude to the friends I made along the way, outside the walls of the office, who made me feel part of a broader and boundless community, both during my time in Trento (Penzuccio, Andrea, Aprilia, Davide, Tina, Goli, Frangè, Claudia, Camilla) and during my time abroad (Ema, Matteo, Margherita, and Alejandro).

A special thanks goes to my family, who agreed to be part of this journey—even without always understanding where, or why, I was heading—and who always supported me unconditionally. Thank you for granting me the privilege to even consider pursuing a PhD. Thanks to Martina, who put up with my strange working hours and the light constantly on during the night. Thanks to Simone, for always being there, in the good days and in the bad ones, for being my safe haven through all these years.

Last but not least, if this thesis saw the light of day, and if I managed to bring this journey to a close, it is also thanks to my thesis reviewers, Prof. Elena Cabrio and Prof. Malvina Nissim, as well as to the rest of the committee members, Dr. Mauro Dragoni and Prof. Massimo Zancanaro.

I am well aware that, although this climb has come to an end, *there's always gonna be another mountain.* And in all honesty, if the next journey is even a fraction as exciting as this one, I already can't wait to start.

Contents

List of Figures

List of Tables

List of Examples

1

Introduction

The self-evident distinction between truth and falsehood is blurring. At least in public opinion, where personal beliefs are shaped more by emotional cues than by objective facts. This erosion of objective truth as a social anchor has led scholars to define a new era: the Post-Truth era (Harsin, 2018). In this era, power rests with those who command the largest audiences, whether political figures or social media influencers capable of mobilizing public sentiment. Knowledge and expertise holders are viewed with suspicion as elitists whose intent, it is claimed, is to undermine the credibility of those able to engage crowds online and offline by leveraging emotion and considerable charisma (Lewandowsky et al., 2017). Truth and falsehood are no longer opposites but rather exist on a blurred spectrum, where facts are accepted or rejected based on their adherence to a set of pre-constituted beliefs rather than their correspondence to reality. In this era of truth extravaganza, misinformation finds fertile ground to take root and proliferate, steadily reshaping the democratic landscape of our societies (Lazer et al., 2018).

Misinformation has become increasingly pervasive in modern society (Lazer et al., 2018), with significant impacts at both societal and individual levels (Adams et al., 2023). At the individual level, misinformation can strongly influence people's beliefs and behaviours in response to false claims (Lewandowsky et al., 2012). Well-known and evident examples include vaccine hesitancy, religious extremism, and disengagement from political participation (Loomba et al., 2021; Booth et al., 2024; Ecker et al., 2024). At the societal level, misinformation undermines trust in media (Wagner and Boczkowski, 2019), erodes public understanding of science, particularly on critical issues like health and climate change (Lewandowsky et al., 2017), destabilizes markets (Petratos, 2021; Kogan et al., 2023), and poses a serious threat to democracy by preventing the electorate from being accurately informed (Kuklinski et al., 2000). The 2024 Global Risks Report[1] identifies misinformation and disinformation as the most severe short-term global risk for fueling polarization, civil unrest, and

[1] https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

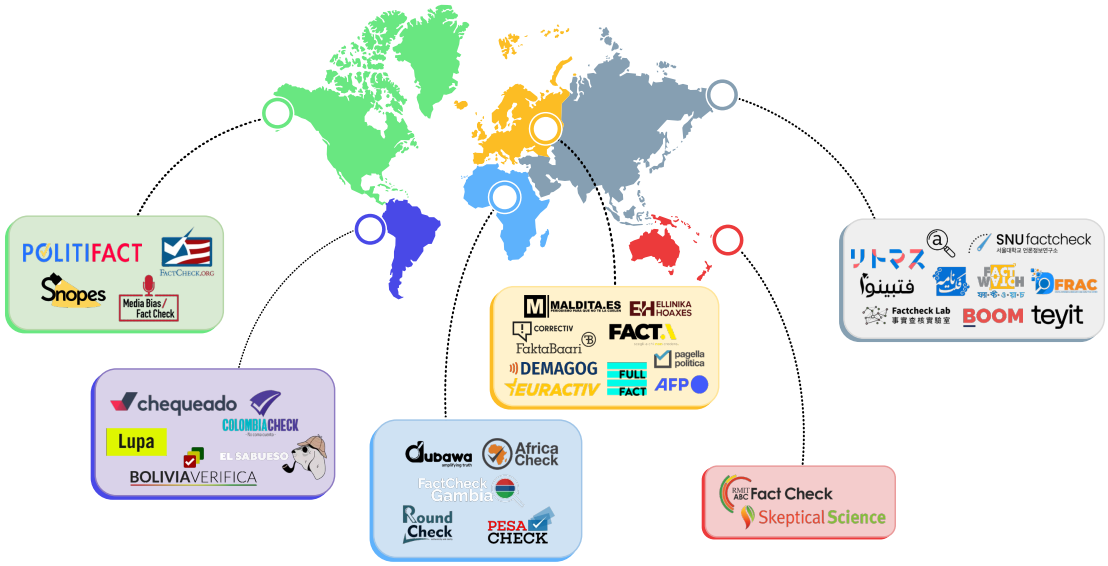


Figure 1.1: Geographical coverage of fact-checking organisations worldwide.

the erosion of democratic rights.

Over the past decade, significant efforts have been made to counter misinformation worldwide (Cazzamatta, 2024; Humprecht, 2019), as evidenced by the proliferation of fact-checking organizations and growing research interest (Lewandowsky and van der Linden, 2021). A graphical overview of several major fact-checking organisations worldwide is presented in Figure 1.1. For a more comprehensive overview of fact-checking organisations worldwide, we suggest the reader consult the list by the Duke Reporters' Lab[2]. Experts from different fields are involved in this arms race, proposing novel solutions and adapting consolidated approaches to the evolving forms that misinformation takes. More generally, two strategies can be distinguished: pre-emptive intervention, namely *prebunking*, and reactive intervention, namely *debunking* (Ecker et al., 2022). Prebunking seeks to help people recognize and resist subsequently encountered misinformation, even if it is novel. Debunking emphasizes responding to specific misinformation, after exposure, to demonstrate why it is false. Although there is not yet a consensus on which strategy is superior, both are considered effective in partially curbing the harm caused by misinformation.

Despite the presence of numerous organizations involved in verifying claims' truthfulness, it is becoming increasingly difficult for fact-checkers to keep up with the relentless growth of false content requiring verification (Adair et al., 2017). This is corroborated by several studies showing that fewer than half of all published news articles have been verified (Lewis et al., 2008; Godler and Reich, 2017). It is therefore clear that manual fact-checking alone is insufficient to efficiently counter the phenomenon of online misinformation. For this reason, Natural Language Processing (NLP) has been proposed as a viable solution for partially automating the fact-checking process. In fact, the manual fact-checking process can be subdivided

[2]https://reporterslab.org/fact-checking

into several mutually dependent and consecutive stages: selecting claims to be examined, gathering sufficient evidence, and producing a verdict that states the degree of truthfulness of the claim along with an explanation of the judgment. Each of these stages can be modelled as NLP tasks already well established in the literature (Vlachos and Riedel, 2014; Kotonya and Toni, 2020a; Guo et al., 2022).

The NLP community has focused primarily on classifying and flagging misleading information. However, classification alone can generate a *backfire effect*, where belief in false claims becomes further entrenched rather than diminished (Lewandowsky et al., 2012). Explaining *why* a claim is classified as true or false offers a more promising solution. While fact-checking articles could serve as a valuable resource toward this end, they prove ineffective on social media platforms for two key reasons: ordinary users are reluctant to click on links to external resources (Glenski et al., 2017, 2020), and these articles are often excessively long, discouraging users from reading them (Pernice et al., 2019). Indeed, effective explanations should be concise, providing only a few key arguments to avoid an *"overkill" backfire effect* (Lombrozo, 2007; Sanna and Schwarz, 2006). The task of automatically generating such explanations has been referred to as justification, explanation, or verdict production/generation. For simplicity, we hereafter use the term "*verdict generation*" to refer to this task.

Early approaches to verdict generation employed rule-based logic and attention token highlighting as forms of explanation for the degree of truthfulness of a given claim (Guo et al., 2022). However, these methods produced verdicts that were difficult to interpret or accessible only to experts. To address readability concerns, summarization-based approaches were then proposed (Atanasova et al., 2020a; Kotonya and Toni, 2020b), which directly extracted knowledge from gold fact-checking articles, representing a significant advance in accessibility. The advent of Large Language Models (LLMs) has since marked a further shift in this landscape, introducing knowledge-driven generative approaches, including retrieval-augmented generation (RAG), that move beyond task-specific extraction towards more flexible and reasoning-capable systems. Unlike summarization-based methods, LLMs are not constrained to a specific downstream task, offering greater versatility in how verdicts are formulated and explained. Despite this potential, systematic studies on the capabilities of current generative LLMs for verdict generation remain scarce. Moreover, while LLMs-based systems for verdict generation offer greater versatility and reasoning capabilities compared to earlier approaches, they largely inherit the same critical limitation: they fail to account for the real-world dynamics of misinformation. Specifically, existing approaches assume ideal conditions that rarely hold in practice, such as the availability of high-quality fact-checking articles, claims expressed in standard formats and languages, and uniform communication styles across platforms. These assumptions render current automated fact-checking systems impractical for integration into professional workflows or deployment at scale.

This thesis addresses these limitations by developing knowledge-driven generative systems for misinformation countering that operate under increasingly realistic conditions. We advance the state of the art along four critical dimensions: (i) adapting to the communication styles of different platforms where misinformation circulates, (ii) expanding geographical and linguistic coverage beyond English-centric

approaches, (iii) operating effectively when high-quality gold knowledge is scarce or unavailable, and (iv) handling different manifestations of misinformation beyond traditional false claims, in particular clickbait. In parallel, this thesis systematically explores the capabilities and boundaries of generative approaches by testing diverse architectural configurations and knowledge integration strategies. By addressing these challenges, this work bridges the gap between academic research and practical deployment, moving automated fact-checking closer to real-world applicability.

1.1 Research Questions

Building on the limitations identified in existing work, particularly the absence of research in automated knowledge-driven generation that takes into account real-world dynamics of misinformation, this thesis investigates the following overarching question: ***How can knowledge-driven generative approaches for countering misinformation be employed to address increasingly realistic scenarios?*** To answer this question, we decompose it into five interconnected steps, each progressively relaxing the idealized assumptions of prior work while advancing toward practical deployment.

RQ1: How can language models leverage reliable knowledge to generate responses to misleading claims?
This foundational question explores whether generative models can effectively exploit external knowledge sources, specifically fact-checking articles, to produce accurate verdicts. While prior work has established the feasibility of knowledge-driven generation in controlled settings, critical gaps remain regarding information extraction and stylistic adaptation. Building on summarization-based approaches, we investigate two sub-questions:

- RQ1.1: How can we extract salient information from fact-checking articles to generate appropriate responses?
- RQ1.2: Can we generate verdicts while adapting their style to match different fact-checking organizations' communication guidelines through knowledge-driven multitask learning?

These questions establish the baseline capabilities required for knowledge-driven verdict generation and examine whether models can adapt to varying journalistic conventions without sacrificing factual accuracy.

RQ2: How can generated verdicts adapt to interlocutors while maintaining empathetic communication?
Adapting to journalistic styles alone proves insufficient when misinformation circulates on social media platforms. Effective verdicts must be empathetic and adapted to the interlocutor's communication style, rather than merely formatted as professional fact-checking articles. This question addresses a critical gap in automated fact-checking: the transition from journalistic discourse to personalized, empathetic communication that accounts for audience characteristics and emotional states. By combining factual accuracy with empathetic framing, we investigate whether such

approaches can reduce defensive reactions and increase the persuasiveness of automated responses in social media contexts.

While the preceding questions advance the quality and adaptability of generated verdicts, they still rely on two strong assumptions: (i) that fact-checking articles always exist for a given claim, and (ii) that such articles are available in the same language as the claim. Therefore, RQ3 and RQ4 will address these challenges.

RQ3: How effective are knowledge-driven approaches for verdict generation in multilingual scenarios?
Often, in realistic scenarios, fact-checks are either unavailable in the claim's language or absent entirely. Given that misinformation is a global phenomenon transcending linguistic boundaries, we must address whether our approaches can operate across languages and handle cross-lingual knowledge transfer. This yields two specific sub-questions:

- RQ3.1: Can multilingual language models counter misinformation regardless of the language, provided that reliable knowledge is available in the same language as the claim?
- RQ3.2: Can multilingual language models respond to misleading claims when available knowledge exists only in a language different from the claim?

By employing prompt-based multilingual approaches, we examine model robustness in both monolingual and cross-lingual settings. However, even successful multilingual approaches still assume the existence of relevant fact-checking articles, an assumption that fails in the majority of real-world cases.

RQ4: How can we generate accurate, faithful, and style-adapted responses when direct, claim-specific fact-checking articles are unavailable?
This question tackles one of the most significant practical barriers to automated fact-checking. In real-world scenarios, the paired claim-knowledge assumption rarely holds: fact-checking articles may exist but remain unlinked to specific claims, or they may not exist at all. We explore RAG approaches to address these challenges through three progressively more realistic sub-questions:

- RQ4.1: How can we generate high-quality verdicts when fact-checking articles exist but are not directly paired with the specific claim?
- RQ4.2: How can we generate high-quality verdicts when no claim-specific fact-checking article exists, relying only on general reliable evidence retrieved from broader knowledge sources?
- RQ4.3: How the claim's communicative format or style (e.g., sensationalized language, vague phrasing) impact overall RAG system performance in these scenarios?

These sub-questions progressively relax the assumption of perfectly paired claim-knowledge data, moving toward scenarios where systems must retrieve and reason over diverse, indirectly related evidence sources.

The preceding research questions assume misinformation takes the form of discrete, verifiable claims that can be directly fact-checked. However, not all misinformation fits this paradigm. Clickbait headlines exemplify this challenge: they mislead readers through sensationalism and curiosity gaps rather than verifiable falsehoods, creating a mismatch between reader expectations and actual article content. Traditional fact-checking systems cannot address this type of misinformation because the deception lies not in false claims but in the gap between headline promises and delivered information. Moreover, the headline itself must be preprocessed to extract verifiable claims before it can be addressed by knowledge-driven generative systems. Addressing this more complex manifestation of misinformation thus demands a broader conception of knowledge-driven countering, motivating the final research question of this thesis.

RQ5: How can knowledge-driven approaches address complex misinformation that cannot be countered through traditional claim-based fact-checking?
This question investigates whether knowledge-driven generation can counter such misinformation through "factuality completion", whereby relevant information is retrieved from the source article to complete the headline and make it factually verifiable. The completed headline can then be processed by knowledge-driven verdict generation systems to produce accurate, informative responses that address the original misleading framing.

Collectively, these research questions represent a systematic progression from foundational capabilities to increasingly realistic and complex scenarios, advancing automated fact-checking from controlled academic settings toward practical, deployable systems capable of operating in the complex realities of real-world misinformation ecosystems.

1.2 Contributions

The following section presents the contributions of this thesis alongside the main related publications.

- An in-depth analysis and benchmark of summarization approaches for verdict generation, along with the proposal of a novel pipeline that combines extractive and abstractive summarization. The analysis spans two fact-checking datasets written in different journalistic styles, evaluating different summarization algorithms and techniques, models, and model configurations.

  ➟ Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264

- The creation of a large-scale, general-domain dataset for verdict generation in social media contexts. Data were created using a human-in-the-loop approach, leveraging LLM capabilities to rapidly generate synthetic data and employing human post-editing to ensure quality. The resulting verdicts align with the informal dynamics of social networks, address emotions present in claims by de-escalating them when needed, and respond with empathy while remaining factually correct and informative. We tested summarization-based models for the joint generation of verdicts and adaptation to user communication styles.

  ➟ Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023a. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics

- The creation of a multilingual dataset covering eight European languages, developed in collaboration with professional EU fact-checkers. Moving beyond summarization-based approaches, we cast verdict generation as a direct prompting task using more powerful LLMs, testing both in-context learning strategies and fine-tuning. We evaluated model capabilities in multilingual scenarios, examining their ability to work simultaneously across different languages and relaxing the assumption that gold knowledge must exist in the same language as the claim to be fact-checked.

  ➟ Daniel Russo, Fariba Sadeghi, Stefano Menini, and Marco Guerini. 2025b. EuroVerdict: A multilingual dataset for verdict generation against misinformation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16617–16634, Vienna, Austria. Association for Computational Linguistics

- An evaluation of retrieval-augmented generation pipelines for verdict creation, testing the impact of claim style on RAG pipeline performance and addressing scenarios where gold knowledge is not paired with the claim to be fact-checked or does not exist at all.

  ➟ Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2025a. Face the facts! evaluating RAG-based pipelines for professional fact-checking. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 846–865, Hanoi, Vietnam. Association for Computational Linguistics

- The development of First-AID, a tool for creating synthetic or semi-synthetic dialogical datasets for knowledge-dependent tasks.

  ➟ Stefano Menini, Daniel Russo, Alessio Palmero Aprosio, and Marco Guerini. 2025. First-AID: the first annotation interface for grounded dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 563–571, Vienna, Austria. Association for Computational Linguistics

- The creation of ClickBaIT, an Italian dataset for clickbait detection and spoiler generation, along with the proposal of a novel task: clickbait neutralization for diminishing the potential harmful impact of clickbait headlines.

  ➟ Daniel Russo, Oscar Araque, and Marco Guerini. 2024. To click it or not to click it: An Italian dataset for neutralising clickbait headlines. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 829–841, Pisa, Italy. CEUR Workshop Proceedings

1.3 Ethical Considerations

This thesis enters the ongoing debate on the promises and perils of Artificial Intelligence, particularly concerning the rapid advancements enabled by large language model (LLM) technologies. Their potential to augment human reasoning, streamline access to information, and support socially beneficial applications has generated considerable optimism across scientific and policy communities. However, as discussed earlier, misinformation poses one of the most severe global risks of our time, threatening democratic processes, societal cohesion, and institutional trust. A common concern is that LLM-based technologies, when coupled with social media platforms and deployed with malicious or destabilizing intent, could undermine these democratic processes. This thesis addresses these concerns by providing resources and methodologies aimed at countering such dynamics while harnessing LLM capabilities for social good.

We recognize that automating fact-checking processes, while essential given the volume and velocity of online misinformation, introduces important ethical considerations. Automatic systems are fallible: detection models may mislabel claims, and generative models can produce factual inaccuracies or hallucinations. Therefore, the systems proposed in this work are not intended to function as standalone tools or replace professional fact-checkers. Instead, we position them as supportive technologies that assist fact-checkers by accelerating verification, enhancing efficiency, and enabling professionals to focus their expertise on the most challenging cases. Human oversight remains essential to ensure accuracy, fairness, and accountability in the fact-checking process.

We also acknowledge that any technological advancement, including those presented here, can potentially be exploited by malicious actors as well. The same building blocks we employ, such as LLMs and knowledge-driven generation techniques, could theoretically be adapted to accomplish opposite goals, such as generating persuasive disinformation or evading detection systems. However, malicious actors will pursue their objectives regardless of community efforts to counter misinformation. Our work contributes to keeping pace with these threats while advancing the state of the art in misinformation countering. Crucially, we rely exclusively on publicly available data and models and publicly release our datasets, code, and fine-tuned models. This commitment to transparency and open science ensures that defensive technologies remain accessible to researchers, fact-checkers, and civil society organizations worldwide, rather than being concentrated in the hands of a few well-resourced institutions.

The datasets created and employed throughout this thesis have been developed with careful attention to ethical considerations. When working with human annotators or fact-checking professionals, we ensured fair compensation and working conditions. For synthetic or semi-synthetic data generation involving LLMs, we implemented human-in-the-loop validation to maintain quality and mitigate potential biases or harmful content. Our multilingual work was conducted in collaboration with professional fact-checkers across European countries, ensuring cultural and linguistic diversity.

Finally, we recognize that the Post-Truth era presents challenges that extend beyond technological solutions. Effective countering of misinformation requires not only accurate detection and explanation systems but also approaches grounded in empathy, cultural awareness, and understanding of human psychology. Throughout this thesis, we have strived to develop systems that are not merely technically proficient but also practically deployable, ethically sound, and aligned with the values of democratic societies. We hope this work contributes to a broader ecosystem of tools and practices that can help restore trust in information and strengthen the resilience of our democratic institutions against misinformation.

1.4 Chapters Outline

The chapters of this thesis follow the same progression as the research questions, gradually addressing increasingly realistic scenarios for knowledge-driven generative approaches to verdict generation. In particular, the scenarios addressed are those already presented at the beginning of this chapter: communication style, linguistic and cultural coverage, knowledge availability, and misinformation typology. Alongside these scenarios, the work presented employs models and architectures of increasing complexity in terms of model size, task formulation, and overall system configuration. It is important to highlight that this work began before the advent of current state-of-the-art LLMs; therefore, some of the technologies proposed and employed may seem outdated. Nevertheless, we strongly believe that the findings, resources, ideas, and concerns raised remain valuable for current and future research, regardless of specific model choices and task formulations.

This introductory chapter is followed by a background chapter (**Chapter 2**), which provides the reader with a comprehensive contextualization of misinformation, drawing from different fields to reveal the multifaceted nature of this phenomenon and to justify the methodological choices made in subsequent chapters. In Chapters 3 and 4, we cast verdict generation as a summarization-based task, focusing not only on producing high-quality verdicts but also on stylistic adaptation. In **Chapter 3**, we test different summarization approaches and evaluate their ability to adapt journalistic style according to the source fact-checking article, using two curated datasets. In **Chapter 4**, the focus shifts to the interlocutor, specifically the user spreading misinformation. The setting moves from a journalistic scenario to a social media one, requiring a shift in communication style and consequently in the types of claims being fact-checked. The style of the verdict must also adapt to this novel scenario. Therefore, the best-performing summarization pipelines from the first study are tested on the joint task of verdict production and style adaptability to the interlocutor's emotional state and communication style. **Chapter 5**

acknowledges that misinformation is not solely an English-centric problem and addresses its broader impact by proposing a novel, manually and professionally curated multilingual resource for verdict generation. To this end, we take a step forward in the technology employed by leveraging a multilingual instruction-based LLM and testing its ability to work with and reason about knowledge regardless of language. This addresses the scenario in which the only available reliable information is not present in the language of the claim. Knowledge may not always be paired with the misleading claim or, more critically, may not yet exist. **Chapter 6** addresses these scenarios by proposing a RAG strategy for verdict generation. Several RAG pipelines are tested not only on their ability to generate verdicts but also on their capacity to produce high-quality verdicts while adopting the appropriate communication style. **Chapter 7** takes a step toward a more challenging task by addressing cases where misinformation is more complex. Specifically, we address the problem of clickbait, where the headline may be harmful depending on the content of the attached article. We tackle this problem by proposing an Italian dataset and three tasks: clickbait detection, spoiler generation, and a novel task we introduce, clickbait neutralization. Finally, **Chapter 8** concludes the thesis by synthesizing the key contributions across all chapters, critically examining the limitations of the proposed approaches, and identifying promising directions for future research.

Table 1.1 provides a comprehensive overview of how each chapter addresses specific realistic dimensions through distinct technological approaches and corresponding research questions.

Realistic Dimension	**Technological Approach**	**RQ**	**Chapter**
Knowkledge Extraction & Communication style	Summarization-based generation with multitask learning	RQ1.1, RQ1.2, RQ2	Chapter 3 Chapter 4
Geographical and linguistic coverage	Multilingual instruction-based LLMs	RQ3.1, RQ3.2	Chapter 5
Knowledge availability & Communication style	Retrieval-augmented generation (RAG)	RQ4.1, RQ4.2, RQ4.3	Chapter 6
Misinformation form (clickbait)	Automatic classification, QA, Instruction-based generation	RQ5	Chapter 7

Table 1.1: Overview of realistic dimensions, technological approaches, research questions, and corresponding chapters of the thesis.

2

Background

In this Chapter, we provide the theoretical background that lays the foundation for the subsequent chapters. The automation of countering strategies against misinformation requires a comprehensive understanding of this multifaceted phenomenon. This chapter is divided into four parts. First, we examine the basic concepts: we review definitions of misinformation and disinformation, existing taxonomies of misinformation types, and discuss the psychological and social mechanisms underlying its spread. Second, we present the main countermeasures currently employed to combat misinformation, focusing on manual fact-checking practices and their inherent challenges and limitations. Third, we introduce the computational approaches developed to partially automate misinformation detection and countering, including knowledge-driven generation techniques. Finally, we present the language modelling foundations underlying verdict generation systems, including neural architectures, decoding techniques, and evaluation of the generations.

2.1 Misinformation and Disinformation

In this section, we examine the relevant theoretical foundations, providing the conceptual framework necessary for understanding knowledge-driven approaches to countering these phenomena. We begin by presenting definitions of misinformation and disinformation, drawing from both philosophical and social science perspectives. We then explore the various types of misinformation through complementary content-based and intent-based taxonomies, examining how false information manifests across different formats and motivations. Finally, we analyse the spreading mechanisms of misinformation, identifying the key actors involved in its dissemination, their motivating factors, and the propagation pathways through which false information circulates in digital ecosystems.

2.1.1 Definitions

Finding a consensus on the definitions of misinformation and disinformation is far from straightforward, particularly as research on these phenomena has expanded across multiple disciplines, each bringing its own conceptual frameworks and terminological preferences.

From a philosophical perspective, several attempts to develop unified accounts of misinformation and disinformation have been proposed (Søe, 2021; Wardle and Derakhshan, 2017). Across this literature, researchers generally agree that three core elements must be considered: *information*, *truth*, and *intent*. However, what ultimately distinguishes misinformation from disinformation depends fundamentally on how "information" itself is understood. According to one influential theory of semantic information (Dretske, 1981), information is inherently true, agent-independent, and fundamental. In this view, misinformation emerges during belief formation: as agents interpret raw information, they may introduce distortions, producing beliefs that no longer carry genuine informational content. For Dretske, misinformation represents "meaning without truth", the outcome of misinterpretation rather than intentional deception. An alternative approach treats information and misinformation as opposites with fixed truth values: information corresponds to the true part of a message, while misinformation corresponds to the false part (Sequoiah-Grayson and Floridi, 2022; Floridi, 2005). Though Floridi shares with Dretske the view of information as objective and agent-independent, he extends these characteristics to misinformation as well, treating both as having determinate truth values. A third perspective moves away from viewing information as inherently true, instead treating it as alethically neutral, that is, a proposition may count as information regardless of whether it is true, false, or lacks a determinate truth value (Fox, 1983; Fetzer, 2004a,b; Fallis, 2014, 2015). Fox treats information and misinformation on par, with misinformation corresponding to the "known-to-be-false" portion of a message. More broadly, within this framework, misinformation can be defined simply as false or misleading information, where intention plays no role in determining its status. By contrast, there is broad agreement that disinformation necessarily involves agents and intentionality. Disinformation arises when someone deliberately seeks to mislead or deceive others (Fetzer, 2004b). As Fetzer succinctly puts it, disinformation is "misinformation with an attitude" (Fetzer, 2004a).

Fetzer's philosophical distinction aligns closely with how these terms are operationalized by practitioners, fact-checkers, policymakers, and scholars in the social and political sciences. In these fields, the boundary between misinformation and disinformation is typically drawn in terms of intent: both involve false information, but disinformation is produced or shared with the purpose of deceiving, whereas misinformation may be accidental or unintentional (Born and Edgington, 2017; Wardle and Derakhshan, 2017; Tucker et al., 2018; Guess and Lyons, 2020). Building on this intent-based distinction, Wardle and Derakhshan (2017) proposed an influential interdisciplinary framework for understanding "information disorder". Their tripartite classification distinguishes:

- *Misinformation*: False information shared without intent to harm
- *Disinformation*: False information deliberately shared to deceive

- *Malinformation*: Genuine information shared with the intent to cause harm, often by exposing material intended to remain private

This framework has been widely adopted in policy circles and by fact-checking organizations (Lazer et al., 2018; Petratos, 2021). It emphasizes two key dimensions for analysis: the authenticity of content (whether information is true or false) and the intent behind its dissemination (whether to inform, mislead, or harm).

Following Wardle and Derakhshan (2017) framework, numerous attempts have been made to systematize the expanding terminology. A commonly adopted strategy focuses on two principal analytical dimensions: (1) the intent to deceive or cause harm, and (2) the authenticity of content. Aïmeur et al. (2023) provide a comprehensive review that synthesizes these frameworks, offering a systematic classification. To these established frameworks must be added consideration of the evolving social media environment. Empowered by generative AI technologies, the landscape has given rise to novel forms of disinformation, including deepfakes and synthetic media, which challenge traditional definitions based solely on intent and authenticity (Ecker et al., 2022).

While philosophical debates continue about the fundamental nature of information and misinformation, there is growing convergence in practical applications around an intent-based distinction that aligns with how these phenomena manifest in real-world contexts. For this thesis, we adopt the widely accepted social science framework: *misinformation* refers to false or misleading information regardless of intent, while *disinformation* specifically denotes false information created or shared with deliberate intent to deceive. This distinction, though not without philosophical complications, provides a pragmatic foundation for developing automated detection and counter-misinformation systems. However, it is important to specify that throughout this work, we predominantly employ the term misinformation as a generalization encompassing both phenomena. This choice reflects the practical reality that automatically detecting intent is intrinsically difficult for automatic systems and remains outside the scope of the studies proposed within this thesis. Therefore, unless explicitly stated, our use of misinformation encompasses false or misleading information irrespective of the underlying intent behind its creation or dissemination.

A related, and equally important, caveat concerns the notion of truth itself, which we have so far treated as a largely unproblematic ingredient of these definitions. Outside of a comparatively small set of sharply factual, checkable claims, truth rarely has the clean boundaries that a binary or even ordinal label presupposes. This tension has long been debated in the fact-checking literature: Uscinski and Butler (2013) argue that journalistic fact-checking rests on a "naïve political epistemology", pointing for instance to the common practice of treating a statement containing multiple component facts as if it were a single fact, or of assessing predictions about future events as though they were already verifiable. Subsequent work has nuanced this critique without dismissing it: Graves (2017) shows, through an ethnographic study of professional fact-checkers, that verification in practice often relies on *factual coherence* with a broader body of evidence rather than on strict, context-free correspondence to reality, while Amazeen (2015) contends that indisputable facts and clearly deceptive claims remain common enough to justify the enterprise despite

its imperfections. The same tension resurfaces, in a more constrained form, when truthfulness must be operationalised for computational systems. Vlachos and Riedel (2014) proposed moving from binary to ordinal veracity labels precisely to better reflect the multi-point scales (e.g. *true*, *mostly-true*, *half-true*) used by journalistic agencies such as PolitiFact, yet Thorne and Vlachos (2018) note that the reasoning behind these finer-grained distinctions is "complex, sometimes inconsistent, and likely to be difficult to capture and express in our models". In practice, this granularity is often reduced again for tractability: Guo et al. (2022) observe that label sets across NLP fact-checking datasets range from 2 to 27 classes, with some works such as Augenstein et al. (2019) leaving them unharmonised and others such as Hanselowski et al. (2019) collapsing them through post-hoc normalisation. As Guo et al. (2022) further note, even a strict true/false separation can be too simplistic, since claims that are technically accurate may still mislead through cherry-picking or selective framing. We therefore want to be explicit that the binary true/false distinction underlying the definitions adopted in this thesis is a necessary simplification rather than a claim about the actual nature of truth: it allows us to operationalise and evaluate countering interventions, but should not be read as a stance on the underlying philosophical question of where the boundaries of truth lie.

2.1.2 Types of Misinformation

Having established the conceptual framework for misinformation and disinformation, we now examine their specific manifestations. The literature proposes various categorizations based on different analytical perspectives. Following Aïmeur et al. (2023), false information can be classified along two complementary dimensions: *content-based* and *intent-based* categorization. These taxonomies are not mutually exclusive but rather provide complementary lenses for understanding the multifaceted nature of false information in digital environments.

Content-Based Classification

Content-based categorization focuses on the format and medium through which false information is disseminated (Parikh and Atrey, 2018):

- **Textual Misinformation**: False information conveyed through written content, including fabricated articles, misleading headlines, and deceptive hyperlinks. This represents the most traditional and prevalent form across online platforms.
- **Multimedia Misinformation**: False or manipulated visual and audio content, including manipulated images, deepfake videos (Chesney and Citron, 2019), fabricated audio, and GAN-generated content (Tolosana et al., 2020). These AI-generated materials can be difficult for common users to distinguish from authentic media.
- **Multimodal Misinformation**: Content combining multiple data types, such as fabricated images paired with misleading text (Nakamura et al., 2020). The coherence between modalities can amplify deceptive effects, as users may treat cross-modal consistency as a credibility signal. Research indicates that video

content tends to be perceived as more convincing than text or audio alone (Hameleers et al., 2020).

Intent-Based Classification

Intent-based categorization examines the purpose and motivation behind the creation and dissemination of false information, providing insights crucial for developing appropriate countermeasures (Wardle and Derakhshan, 2017; Aïmeur et al., 2023). The principal categories include:

- **Clickbait**: Misleading headlines or thumbnails designed to maximize click-through rates and website traffic, primarily for advertising revenue (Zannettou et al., 2019; Hagen et al., 2022). While deceptive, clickbait is considered less harmful since the full content usually reveal the headline's misleading nature.

- **Hoax**: Intentionally fabricated information presented as factual to deceive audiences (Kumar et al., 2016). Hoaxes often involve sensational claims about events or individuals, exemplified by false celebrity death reports or COVID-19 health misinformation.

- **Rumor**: Unverified claims disseminated without supporting evidence (Zubiaga et al., 2018). Unlike other categories, rumors are not necessarily false and may prove true or remain unresolved. They capitalize on uncertainty and spread rapidly through social networks.

- **Satire**: Content using irony and humor to critique social, political, or cultural issues (Rubin et al., 2016). While potentially presenting factual inaccuracies, satire's intent is to entertain or expose problematic behavior rather than deceive. However, audiences may fail to recognize satirical intent, leading to unintentional misinformation spread. Some platforms exploit satirical disclaimers to shield deliberately false content from accountability.

- **Propaganda**: Information created by political entities to manipulate public opinion and advance specific agendas (Da San Martino et al., 2019). As a state-sponsored activity, propaganda can significantly influence democratic processes. Modern manifestations include astroturfing, covert manipulation creating false impressions of grassroots support (Peng et al., 2016).

- **Framing**: Strategic emphasis or concealment of reality aspects to guide interpretation while obscuring truth (Entman, 1993). Framing operates through selective presentation rather than complete fabrication, exploiting authentic information while fundamentally misrepresenting its meaning.

- **Conspiracy Theories**: Beliefs attributing significant events to secret plots by powerful actors (Douglas et al., 2019). Associated with psychological, political, and social factors, conspiracy theories have become prevalent in contemporary democracies, with serious public health consequences highlighted during the COVID-19 pandemic (Uscinski et al., 2020).

These categories are not mutually exclusive; false information frequently exhibits characteristics of multiple types simultaneously (Zannettou et al., 2019). For example, a rumor may employ clickbait techniques to maximize reach, while propaganda often represents a specific instance of framing. Conspiracy theories may be disseminated through fabricated multimedia content. This overlap between misinformation categories might complicates detection efforts and underscores the need for multifaceted approaches to identifying and mitigating false information across different types and formats.

2.1.3 Spreading

Having examined the definitions, types, and characteristics of misinformation, we now turn to understanding how false information propagates through digital ecosystems and who the key actors are in its dissemination. Understanding the spreading mechanisms and agents involved is crucial for developing effective countermeasures, as different types of misinformation involve distinct actors with varying motivations and employ different propagation strategies.

Actors and Motivations

Identifying the agents involved in misinformation dissemination and their modus operandi presents significant challenges, as those who aim to deceive online users typically conceal their identities (Guess and Lyons, 2020). However, some investigations have successfully unmasked groups responsible for contaminating the media ecosystem with false content. Notable examples from studies on the 2016 United States presidential elections include a group of teenagers in Veles, Republic of North Macedonia, who registered approximately 100 pro-Trump fake news websites (Kirby, 2016; Subramanian, 2017), and Russia's Internet Research Agency (IRA), which disseminated false content through fabricated social media accounts on Facebook and Twitter (Bastos and Farkas, 2019). In 2018, Twitter identified 3,841 accounts managed by the IRA.[1]

As the misinformation phenomenon has evolved, so too have the agents involved in spreading harmful content. Wardle and Derakhshan (2017) delineates a comprehensive typology distinguishing between *official actors* (intelligence services, political parties, news organizations) and *unofficial actors* (groups of citizens mobilized around specific issues). Disinformation agents operate across various organizational structures, from individuals working alone to tightly-organized longstanding organizations (such as PR firms or lobbying groups) to impromptu groups organized around common interests.

Four primary motivating factors drive these actors:

- **Financial**: Profiting through advertising revenue. For instance, the Veles teenagers earned up to $8,000 per month in a country where the average monthly wage was approximately $400. Producers of false content exploit the chaotic social media ecosystem by creating content that capitalizes on user limitations and emotional needs to maximize engagement and, consequently, profit (Guess and Lyons, 2020).

[1] https://blog.twitter.com/en_us/topics/company/2018/2016-election-update

- **Political**: Influencing public opinion, discrediting political candidates, or shaping electoral outcomes.
- **Social**: Establishing connections with specific online or offline groups.
- **Psychological**: Seeking prestige, validation, or reinforcement of existing beliefs.

Different agents target different audiences, ranging from internal organizational mailing lists to specific social groups based on socioeconomic characteristics to entire societies.

Propagation Mechanisms

Understanding the modalities through which false and misleading content spreads online is essential for developing effective interventions. Most research on misinformation propagation has focused on Twitter due to its API accessibility compared to other platforms like Facebook. Multiple studies have identified *social bots* as significant amplifiers of disinformation during the early stages of fake news dissemination (Bessi and Ferrara, 2016; Ferrara, 2017; Shao et al., 2018). Ferrara et al. (2016) define social bots as covert computer algorithms that automatically produce content and interact with humans on social media, attempting to emulate and possibly alter their behaviour.

Another crucial propagation mechanism involves *breaking news* sites and accounts, which deliberately mimic legitimate information sources through careful aesthetic and stylistic design (Andrews et al., 2016). Twitter users often trust these accounts due to their resemblance to authentic news organizations.

Beyond technological factors, a complex interplay of cognitive, social, and algorithmic biases contributes to increased user vulnerability and exposure to false content (Shao et al., 2018). Human attention limitations cannot adequately filter the overwhelming flow of online misinformation, which gains popularity as rapidly as legitimate news (Qiu et al., 2017). This scenario is exacerbated by users' tendency to aggregate in homogeneous online groups sharing common beliefs, creating *echo chambers*. Cinelli et al. (2021) define echo chambers as environments in which users' opinions, political leanings, or beliefs about a topic become reinforced through repeated interactions with peers or sources having similar tendencies and attitudes. Echo chambers amplify the circulation of false or misleading content that aligns with community beliefs (Shin et al., 2017; Guess and Lyons, 2020).

These cognitive and social biases are further reinforced by social media algorithms designed to provide personalized content. Research demonstrates that these algorithms tend to promote content based on popularity rather than quality, prioritizing engagement over accuracy (Ciampaglia et al., 2018).

Synthesis: Types, Actors, and Propagation

Table 2.1 synthesizes the relationships between misinformation types, their characteristics, typical actors, and propagation patterns discussed in the previous subsections. This comparative framework illustrates how different forms of misinformation combine varying levels of deceptive intent, propagation speeds, potential harm, and

actor motivations. The most dangerous forms of misinformation combine high deceptive intent, rapid propagation, significant potential harm, and complex hidden motivations. Propaganda, rumors, and conspiracy theories emerge as particularly problematic due to their widespread societal impact, rapid dissemination, and the organized nature of their typical actors (Aïmeur et al., 2023).

Type	Veracity	Intent to Deceive	Propagation Speed	Potential Harm	Primary Goal	Typical Actors
Clickbait	Misleading/False	High	Slow	Low	Profit	Publishers
Hoax	False	High	Fast	Medium	Various	Individuals
Rumor	Uncertain	Medium	Fast	High	Information	Public
Satire	False (intentional)	Low	Slow	Low	Entertainment	Comedians
Propaganda	Mixed	High	Fast	High	Political	States/Parties
Framing	Selective truth	High	Fast	Medium	Manipulation	Media/Political
Conspiracy	False narrative	High	Fast	High	Ideological	Communities

Table 2.1: Comparative analysis of intent-based misinformation types, synthesizing relationships between content characteristics, actor motivations, and propagation patterns. Adapted from Aïmeur et al. (2023); Wardle and Derakhshan (2017); Zannettou et al. (2019).

2.2 Countermeasures

The pervasive spread of misinformation demands effective intervention strategies at multiple stages of its lifecycle. Over the past decade, considerable research effort has been devoted to developing and evaluating approaches to counter misinformation (Lewandowsky et al., 2020; Ecker et al., 2022). This section provides an overview of the primary countermeasures that have emerged from both research and practice, organized according to their temporal relationship to misinformation exposure and their theoretical foundations. We begin by examining *prebunking* strategies, which operate pre-emptively to build cognitive resilience before individuals encounter false or misleading claims. We then turn to *debunking and fact-checking*, the reactive interventions that respond to specific instances of misinformation after exposure. Finally, we synthesize *insights from cognitive science* regarding the efficacy of these interventions. Drawing on psychological and neuroscientific evidence, we explore why people form false beliefs in the first place, examine the conditions under which corrections successfully revise these beliefs, and consider how emotional content and social contexts influence correction effectiveness.

Together, these subsections provide a comprehensive understanding of current countermeasures to misinformation, establishing the theoretical and empirical foundation for the knowledge-driven generative approaches developed in subsequent chapters of this thesis.

2.2.1 Prebunking

Prebunking encompasses pre-emptive interventions designed to build cognitive resilience before individuals encounter misinformation (Ecker et al., 2022). Unlike reactive debunking that addresses false claims after exposure, prebunking aims to equip people with the tools to recognize and resist misleading content proactively.

Research has identified a spectrum of prebunking strategies that vary in sophistication and theoretical foundation.

At the simpler end of this spectrum are interventions that provide factually correct information in advance (Dai et al., 2021; Swire-Thompson et al., 2021), issue pre-emptive corrections (Brashier et al., 2021; Gordon et al., 2019), or display generic warnings about misinformation (Ecker et al., 2010; Clayton et al., 2020). These approaches operate on the assumption that forewarning or pre-arming individuals with accurate information reduces their susceptibility to subsequent false claims.

More theoretically grounded prebunking interventions draw on inoculation theory, which applies the logic of biological vaccination to informational contexts (Compton et al., 2021; Lewandowsky and van der Linden, 2021). Just as vaccines expose individuals to weakened pathogens to build immunity, inoculation interventions expose people to weakened forms of persuasive arguments to stimulate critical resistance. These interventions typically combine two core components: first, they warn recipients about the threat of being misled; second, they expose and explain the manipulative techniques commonly used in misinformation (Ecker et al., 2022). By understanding these tactics, individuals develop generalizable cognitive defenses that transfer across topics (Parker et al., 2012).

A related approach focuses on enhancing information literacy and media literacy, skills that enable individuals to effectively find, evaluate, and use information. Improved information literacy has been associated with better detection of false news (Jones-Jang et al., 2021) and reduced sharing of misinformation (Khan and Idris, 2019). One particularly effective technique is lateral reading, where individuals consult external sources to verify the credibility and plausibility of claims or their sources (Amazeen and Krishna, 2020; Breakstone et al., 2021). However, evidence regarding the effectiveness of literacy interventions remains mixed (Amazeen and Bucy, 2019; Guess et al., 2020). Passive literacy interventions have shown limited success, while more active approaches, particularly game-based interventions that simulate misinformation environments and allow users to practice identifying manipulative techniques, have demonstrated greater promise (Roozenbeek and Van Der Linden, 2019; Roozenbeek and van der Linden, 2020; Katsaounidou et al., 2019).

Despite its advantages, prebunking faces inherent limitations. Its strength lies in building general resistance across diverse misinformation types, but this generalizability comes at a cost: prebunking is less effective when specific, circulating pieces of misinformation require targeted responses (Ecker et al., 2022). Consequently, prebunking and debunking are best understood as complementary strategies within a comprehensive intervention framework, with prebunking providing broad-spectrum protection and debunking addressing specific false claims.

2.2.2 Debunking and Fact-checking

While prebunking operates proactively, debunking and fact-checking constitute reactive interventions that respond to specific instances of misinformation after exposure (Ecker et al., 2022). Although these terms are often used interchangeably, they refer to distinct but related practices. **Fact-checking** encompasses the systematic verification of factual claims in texts or speeches to determine their accuracy (Graves and Cherubini, 2016). **Debunking** has a broader scope, aiming to expose false-

hoods or demonstrate that information is less significant, credible, or true than it has been made to appear (Pamment and Kimber, 2021). In practice, professional fact-checking organizations engage in both activities: verifying factual accuracy and debunking misleading narratives.

Over the past decade, numerous public and private organizations have emerged with the mandate of verifying controversial and misleading claims (Graves and Cherubini, 2016). As illustrated in Figure 1.1, fact-checking organizations now operate across multiple continents, though coverage remains uneven globally. These organizations typically adhere to shared methodological principles and ethical standards, often formalized through networks such as the International Fact-Checking Network (IFCN), which establishes codes of practice for transparency, methodology, and independence.

Despite variations in organizational structure and focus, most fact-checking organizations follow a similar process methodology (Graves, 2017). This process can be generalized into four main stages:

1. **Claim selection.** The fact-checking process begins with identifying claims worthy of verification. Claims must contain verifiable facts of public interest, and organizations typically prioritize those with the potential to significantly impact public discourse or people's lives. Selection criteria often balance timeliness, reach (how widely the claim has spread), and potential harm.

2. **Evidence retrieval.** Determining a claim's truthfulness requires gathering substantial, reliable evidence. Fact-checkers consult academic research, government reports, expert sources, and primary documents. Organizations often attempt to contact claimants directly to understand the basis for their statements and, when necessary, seek expert consultation in specialized domains.

3. **Fact-checking article writing.** The gathered evidence is synthesized into an article designed to provide readers with comprehensive, reliable information necessary to understand the claim and form an informed opinion. These articles typically present evidence transparently, acknowledge uncertainties where they exist, and explain the reasoning process.

4. **Verdict generation.** Finally, a judgment regarding the claim's truthfulness is rendered. Different organizations employ varying formats for presenting verdicts. Some use graded scales that reflect degrees of accuracy (ranging from completely true to completely false, often with intermediate categories), while others integrate verdicts into narrative explanations without rigid categorical labels. Regardless of format, verdicts are typically accompanied by summaries of the justifications that led to the conclusion. The diversity in verdict presentation styles reflects different organizational philosophies about how best to communicate nuanced assessments to diverse audiences.

The differences in verification workflows and verdict presentation formats across organizations present both challenges and opportunities for automation. Chapter 3 examines how these stylistic variations can be leveraged in automated verdict generation systems, using data from organizations with distinct presentation styles to

develop adaptable generation approaches. Understanding these professional practices is essential for designing automated systems that can complement rather than replace human fact-checkers.

2.2.3 Efficacy: Insights from Cognitive Science

Understanding the efficacy of fact-checking interventions requires examining the psychological mechanisms underlying misinformation belief and correction. Drawing on cognitive science research, this section addresses three fundamental questions: (i) why people believe misinformation, (ii) whether corrections effectively revise false beliefs, and (iii) how these processes operate in emotionally charged social media contexts.

False beliefs arise through the same cognitive, social, and affective mechanisms that establish accurate beliefs (Marsh et al., 2016; Rapp, 2016). People are naturally biased to believe information (Pantazi et al., 2018) and rely on intuitive heuristics rather than deliberative reasoning (Brashier and Marsh, 2020).

A particularly robust mechanism is the *illusory truth effect*: repeated exposure to a claim increases its perceived believability (Unkelbach et al., 2019). This occurs because people use peripheral cues such as familiarity, processing fluency, and coherence as truth signals (Begg et al., 1992; Unkelbach and Rom, 2017). These effects persist for months (Brown and Nix, 1996), regardless of cognitive ability (De keersmaecker et al., 2020), and despite contradictory information from credible sources (Unkelbach and Greifeneder, 2018).

Beyond cognitive shortcuts, worldview congruence strongly influences belief formation. People more readily accept information aligning with their existing views (Pennycook and Rand, 2021), though difficulties discerning truth from falsehood can also stem from intuitive thinking independent of ideological alignment (Pennycook and Rand, 2019). Source credibility also matters: messages from perceived experts or in-group members are judged more trustworthy (Nadarevic et al., 2020), yet people frequently forget or overlook source information (Mitchell and Johnson, 2009), which is problematic given that small numbers of accounts spread disproportionate misinformation (Grinberg et al., 2019).

Research demonstrates that direct corrections are effective in reducing misinformation reliance, though rarely eliminating it entirely (Chan et al., 2017; Walter and Murphy, 2018). The most effective corrections provide detailed factual refutations with alternative explanations rather than simple retractions (Johnson and Seifert, 1994; Ecker et al., 2020; Lewandowsky et al., 2012). Indeed, merely labelling a claim as true or false can trigger the *backfire effect*, in which the denial itself reinforces belief in the false information, as the correction fails to provide an alternative explanation to replace it (Lewandowsky et al., 2012). Providing an alternative causal explanation fills the resulting gap in the recipient's mental model, thereby reducing the backfire effect and yielding a more effective correction. For instance, correcting the false claim that a vaccine causes autism is more effective when accompanied by an explanation of the biological mechanism by which the immune system safely processes vaccine components, rather than by a simple denial of the alleged link. Moreover, effective explanations should be simple, and only a few arguments must be provided in order to avoid an *"overkill" backfire effect* (Lombrozo, 2007; Sanna

and Schwarz, 2006).

Despite these reassurances, misinformation can continue influencing reasoning even after correction acceptance, a phenomenon termed the *continued influence effect* (CIE) (Johnson and Seifert, 1994; Chan et al., 2017). The CIE arises because misinformation and corrections coexist in memory rather than the former being erased (Shtulman and Valcarcel, 2012). Two theoretical accounts explain this persistence: corrections may fail to integrate with misinformation in neural memory networks (Ecker et al., 2010; Kendeou et al., 2014), or misinformation may be automatically retrieved without concurrent recollection of the correction (Ecker et al., 2011; Yonelinas, 2002). Both perspectives support the superiority of detailed refutations, which either strengthen correct information activation or provide more retrieval cues (Lewandowsky et al., 2012; Kendeou et al., 2019).

Corrections on social media can reduce false beliefs among both targets and observers, a process termed *observational correction* (Vraga and Bode, 2017). Effective social media corrections are content-specific rather than generic (Clayton et al., 2020), can originate from automatic systems, expert organizations, or other users (Bode and Vraga, 2015; Vraga and Bode, 2017; Bode and Vraga, 2018), and should be delivered quickly (Vraga and Bode, 2020). The phenomenon in which common users directly reply to misleading content online is known as "social correction" (Ma et al., 2023). Within this context, gentle accuracy nudges may be preferable to public corrections to avoid ostracizing individuals (Mosleh et al., 2021; Pennycook et al., 2021).

Emotion significantly influences correction processing. Misinformation conveying fear or anger may particularly evoke continued influence (Cobb et al., 2013; Thorson, 2016), yet corrections can benefit from *emotional recalibration*. When misinformation downplays risks, corrections providing accurate assessments operate partly through emotional impacts (Van der Meer and Jin, 2020), while countering fear-inducing disinformation benefits from reducing emotional arousal (Featherstone and Zhang, 2020).

In sum, fact-checking interventions effectively reduce misinformation belief and influence when designed with attention to underlying cognitive, social, and affective mechanisms (Ecker et al., 2022). For automated verdict generation, this evidence underscores the importance of providing detailed explanatory corrections rather than simple classifications, supporting the knowledge-driven approaches developed in this thesis.

2.3 Automated Fact-checking

The scalability challenges inherent in manual fact-checking have become increasingly apparent as the volume of misinformation continues to grow exponentially. Despite the efforts of numerous organizations dedicated to verifying claim truthfulness, fact-checkers struggle to keep pace with the relentless proliferation of false information requiring debunking (Adair et al., 2017). Empirical evidence underscores this limitation: studies have shown that less than half of all published articles undergo verification (Lewis et al., 2008; Godler and Reich, 2017). Given these constraints, manual fact-checking alone proves insufficient to effectively counter the phenomenon of online misinformation. These practical limitations have motivated research into

automated approaches, particularly within the domain of computational journalism (Cohen et al., 2011; Graves, 2018).

Graves (2018) identifies three principal objectives that guide research in automated fact-checking: the identification of false or misleading claims circulating online; the authoritative verification of selected claims; and the immediate correction of false statements to mitigate their potential for causing harm. These objectives reflect both the technical capabilities required and the broader societal goals that automated systems must serve.

2.3.1 The Automated-Fact-Checking pipeline

Natural Language Processing (NLP) offers a promising framework for partially automating the fact-checking process. The foundational insight, articulated by Vlachos and Riedel (2014), is that the principal stages of fact-checking can be effectively modeled as NLP tasks. This observation has catalyzed diverse research efforts over the past decade, each targeting one or more tasks of the fact-checking workflow. The NLP framework for automated fact-checking comprises three main stages: claim detection, evidence retrieval, and claim verification (Figure 2.1). The final stage, claim verification, can be further decomposed into two distinct subtasks: verdict prediction and justification production (Kotonya and Toni, 2020a; Guo et al., 2022).

Figure 2.1: NLP framework for Automated Fact-checking processes

Claim Detection

The first stage, claim detection, addresses the problem of identifying which claims warrant verification. This selection process typically relies on the concept of check-worthiness, which Hassan et al. (2015) define as the degree to which the general public would benefit from knowing the truth about a particular claim. The task is commonly formulated either as a binary classification task or as an importance ranking task(Atanasova et al., 2018; Barrón-Cedeño et al., 2020). However, this formulation introduces inherent challenges. The concept of check-worthiness is fundamentally subjective, varying across temporal contexts and potentially embedding systematic biases that disadvantage marginalized communities (Guo et al., 2022). These limitations raise important questions about whose perspectives shape the criteria for what constitutes important information to verify.

Evidence Retrieval

The second stage, evidence retrieval, involves locating relevant information to support or refute a claim. This evidence may take various forms, including text, tables,

images, and other structured or unstructured data sources. Researchers have explored multiple approaches to this challenge. Traditional strategies employ search APIs such as Wikipedia Search or Google Search, Lucene indexes, or named-entity linking techniques (Thorne et al., 2018b). More recently, neural retrieval models have demonstrated strong performance on this task (Maillard et al., 2021). An alternative paradigm, proposed by Fan et al. (2020), frames evidence retrieval in terms of question generation and question answering. A central challenge in this stage concerns balancing the degree of automation with the need for trustworthiness assessment of retrieved information (Guo et al., 2022). Not all sources are equally reliable, and automated systems must navigate this heterogeneity effectively. Sources range from professionally curated fact-checking outlets, which adhere to rigorous verification standards, to conspiracy theory websites, which systematically promote unsubstantiated claims, posing a significant challenge for automated credibility assessment.

Verdict Prediction and Justification Production

The final stage encompasses two complementary tasks: determining whether a claim is true or false, and providing a justification for that determination. Verdict prediction is typically approached as a classification problem. Some systems employ binary classification schemes (Nakashole and Mitchell, 2014; Potthast et al., 2018b; Popat et al., 2018), while others adopt multi-class frameworks that capture gradations of truthfulness (Wang, 2017). Augenstein et al. (2019) advanced this line of research by proposing a multitask learning framework that simultaneously ranks evidence pages and predicts claim veracity, leveraging the synergies between these related objectives.

Justification production presents distinct challenges. While verdict prediction focuses on determining veracity, justification production requires explaining that determination in a manner that is both accurate and persuasive. Researchers have pursued several methodological approaches. Rule-based systems exploit formal reasoning frameworks, leveraging Horn rules and knowledge bases to derive structured explanations (Gad-Elrab et al., 2019; Ahmadi et al., 2019), while deep learning approaches employ attention mechanisms to identify salient evidence (Popat et al., 2018; Yang et al., 2019; Shu et al., 2019; Lu and Li, 2020). However, the most promising direction appears to be framing justification production as a summarization task (Kotonya and Toni, 2020a). Within this paradigm, explanations can be generated through two primary strategies: extractive approaches, which select key sentences from source articles (Atanasova et al., 2020b), and abstractive approaches, which generate novel explanatory text (Kotonya and Toni, 2020b). Each approach offers distinct advantages in terms of faithfulness, fluency, and informativeness.

2.3.2 Verdict Generation

Prior research on verdict generation has explored multiple methodological frameworks, ranging from symbolic reasoning systems to neural architectures. This section reviews the principal approaches that have shaped the field, highlighting both their contributions and their limitations.

Rule-based Approaches

Early attempts at justification production employed symbolic reasoning methods, specifically leveraging Horn rules and knowledge graphs to extract explanations (Gad-Elrab et al., 2019; Ahmadi et al., 2019). In this paradigm, rules are mined from pre-existing knowledge bases such as DBpedia (Auer et al., 2007), within which claims are represented as subject-predicate-object triplets (Figure 2.2). These systems offer the advantage of transparency, as the logical derivations leading to verdicts can be explicitly traced and verified by domain experts.

```
FALSE : almaMater(Michael White, UT Austin)
 ←  employer(Michael White, UT Austin)
 ←  occupation(Michael White, UT Austin)
 ←  almaMater(Michael White, Abilene
     Christian Univ.), almaMater(Michael
     White, Yale Divinity School)
```

(a) (Gad-Elrab et al., 2019)

Fact candidates	Explanations
countryWon(guatemala, nobel)	*isCitizenOf(miguel_asturias, guatemala)*, *source* = TEXT, *hasWonPrize(miguel_asturias, nobel)*, *source* = KG
influences(s_fitzgerald, r_yates)	*wrote(s_fitzgerald, the_great_gatsby)*, *source* = KG *read(r_yates, the_great_gatsby)*, *source* = TEXT, text="Ya

(b) (Ahmadi et al., 2019)

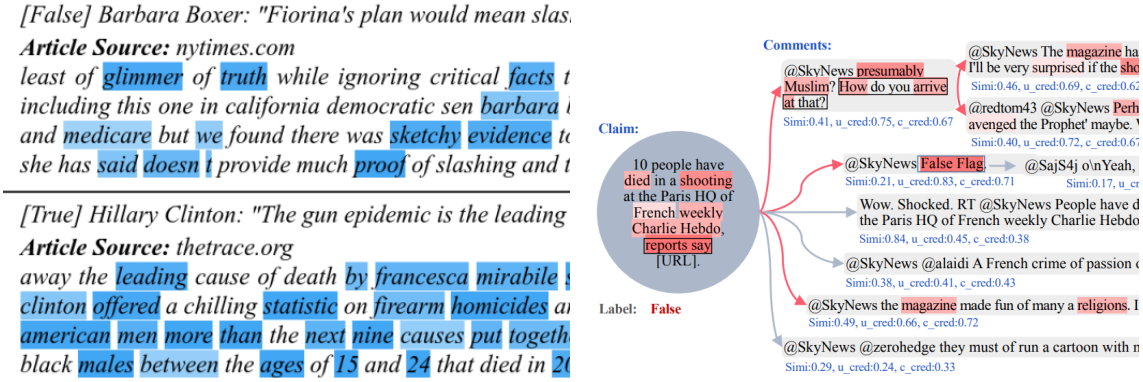
Figure 2.2: An example of rule-based explanations

However, rule-based approaches face significant structural limitations. First, they can only verify claims that can be adequately represented as triplets, which excludes many natural language claims that involve complex temporal, causal, or modal relationships. Second, their coverage is constrained by the information present within the knowledge base itself (Ahmadi et al., 2019; Kotonya and Toni, 2020a). Even well-curated knowledge bases suffer from incompleteness, meaning that many factually verifiable claims may lack the necessary supporting information within the system. These constraints limit the applicability of purely rule-based systems for general-domain fact-checking.

Attention-based Approaches

The advent of deep learning introduced attention mechanisms as an alternative approach to explanation generation. Popat et al. (2018) proposed a bidirectional LSTM architecture designed to assess claim credibility against a corpus of articles. Beyond producing credibility scores, the model generates annotated versions of source articles in which tokens receiving higher attention weights are highlighted as explanations (Figure 2.3a). This approach treats attention weights as a proxy for the relevance of evidence, under the assumption that tokens receiving greater attention from the model correspond to information that was influential in the verification decision.

Subsequent work has explored variations on this theme. Co-attention mechanisms, which jointly attend to both claims and evidence, have been employed to identify salient excerpts (Shu et al., 2019), while self-attention architectures have been applied to the same objective (Yang et al., 2019). Lu and Li (2020) extended this approach to more realistic scenarios by incorporating social media data, specifically from Twitter, using co-attention weights to both assess truthfulness and generate explanations. Wu et al. (2020) proposed a hybrid architecture combining co-attention weights with decision trees, enabling the system not only to highlight evidential tokens but also to provide structured explanations of how these tokens influenced the final verdict (Figure 2.3b).



(a) (Popat et al., 2018) (b) (Wu et al., 2020)

Figure 2.3: An example of attention-based explanations

Despite these advances, recent empirical and theoretical work has raised fundamental questions about the validity of attention-based explanations (Jain and Wallace, 2019). Studies have demonstrated that tokens with high attention weights can sometimes be removed without affecting model predictions, while removal of tokens with low attention weights may dramatically alter model behavior. This phenomenon undermines the faithfulness of attention-based explanations, which refers to how accurately an explanation reflects the model's actual reasoning process (Jacovi and Goldberg, 2020). Additionally, attention-based explanations present readability challenges. Highlighted tokens, even when faithfully representing model attention, may not be intuitively interpretable by users without technical expertise (Guo et al., 2022). These limitations have motivated research into alternative approaches that can provide more reliable and comprehensible explanations.

Summarization-based Approaches

Framing justification production as a summarization task has emerged as a particularly promising direction (Kotonya and Toni, 2020a). This approach leverages the observation that fact-checking articles, written by professional journalists, already contain well-structured explanations that justify verdicts. The challenge then becomes one of selecting or generating similar explanatory text automatically. Within this paradigm, two principal studies have established foundational methodologies: Atanasova et al. (2020b) for extractive summarization and Kotonya and Toni (2020b) for abstractive summarization. Both approaches build upon the architectural innovations presented in Liu and Lapata (2019), which demonstrated effective methods for adapting BERT (Devlin et al., 2019) to summarization tasks through the BertSum framework.

2.4 Language Models for Verdict Generation

This section provides the technical foundations necessary to understand the language model approaches employed for verdict generation throughout this thesis. We begin by introducing language models and the transformer architecture, which form the backbone of modern natural language processing systems. We then examine decoding mechanisms, which govern how language models generate text by selecting to-

kens at each step of the generation process. Subsequently, we discuss automatic text summarisation, as verdict generation is initially formulated as a summarisation task within the present thesis. We then address retrieval augmented generation, an approach that enhances language model outputs by incorporating external knowledge sources to improve factual accuracy and reduce hallucinations, i.e., the generation of factually incorrect information (Huang et al., 2025). Finally, we present evaluation methodologies for assessing the quality of generated text, including both automatic metrics and human evaluation frameworks. Together, these components provide the reader with the technical knowledge required to engage with the methodological choices and experimental designs presented in the following chapters.

2.4.1 The Transformer Architecture

A drastic breakthrough in natural language processing occurred in 2017 with the advent of Transformer models (Vaswani et al., 2017), which rapidly became dominant in the field. Transformers are based on attention mechanisms only, specifically on *self-attention* (also known as *intra-attention*). Similar to RNNs, Transformers process sequential data; however, unlike RNNs, the entire input is processed all at once and not word by word. The attention mechanism makes Transformers state-of-the-art models for three main reasons: (1) it allows the model to drastically reduce the total computational complexity per layer; indeed, self-attention layers are faster than the recurrent layers employed in RNNs and LSTMs; (2) the sequential nature of RNNs and LSTMs hinders parallel computations: with self-attention this limitation is overcome as it increases the amount of computation that can be parallelized; (3) self-attention layers drastically reduce maximum path length between long-range dependencies in the network, allowing the model to learn long-range dependencies more easily and efficiently.

The original Transformer model (Vaswani et al., 2017) employed an encoder-decoder architecture. The encoder comprises six identical layers stacked one on top of the other. Each layer is composed of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Each encoder layer extracts salient features from the input and generates encodings containing this information. Encodings are iteratively produced and passed on to the following encoding layer. Similarly, the decoder consists of a stack of six identical layers. Along with the two aforementioned sub-layers, each decoder layer contains a multi-head attention module over the output of the encoder stack of layers. In order to avoid leftward information flow, the self-attention mechanism is masked within the decoder layers. The function of the decoder layers is the opposite of the encoder ones: they take as input the encodings and use the related contextual information to generate the output textual sequence. Both in the encoder and the decoder, around each sub-layer, residual connections followed by layer normalisation are employed.

Throughout the years, different language models based on the original Transformer (Vaswani et al., 2017) have been deployed. Two main architectural variants emerged: encoder-decoder models and decoder-only models. Encoder-decoder models, such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020), maintain the full transformer architecture and are particularly effective for tasks that require understanding input context and generating coherent output, such as summarization

and translation. Decoder-only models, such as the GPT family, utilize only the decoder component with causal masking, making them well-suited for autoregressive text generation tasks.

Pre-training the entire model using large amounts of data has become a standard procedure. The goal of pre-training is to allow the model to acquire general-purpose linguistic skills and knowledge that can later be transferred through fine-tuning to downstream tasks, such as summarization. During fine-tuning, the weights of some of the model's layers, usually the top layers, are adjusted so that the model is able to acquire abilities toward a specific task. This procedure of leveraging existing general knowledge to specialize a model toward a specific activity is known as transfer learning.

The evolution of transformer-based models has led to the development of Large Language Models (LLMs), which are characterized by their massive scale (billions or even trillions of parameters) and extensive pre-training on diverse text corpora. Contemporary LLMs, such as the Llama family employed in some of the chapters of this thesis, demonstrate remarkable capabilities in few-shot and zero-shot learning, enabling them to perform tasks with minimal task-specific fine-tuning through instruction-following and in-context learning. These models have shown particular promise in knowledge-intensive tasks, though they remain susceptible to hallucinations.

2.4.2 Decoding Mechanisms

In autoregressive language generation, the probability distribution of a word sequence is decomposed into the product of conditional next-word distributions:

$$P(w_{1:T} \mid W_0) = \prod_{t=1}^{T} P(w_t \mid w_{1:t-1}, W_0) \tag{2.1}$$

where W_0 represents the initial context and $w_{1:0} = \emptyset$. The sequence length T is typically determined dynamically when an end-of-sequence (EOS) token is generated. The selection of the next token at each step is governed by a *decoding mechanism*, which plays a crucial role in determining the quality, diversity, and coherence of the generated text.

Decoding strategies can be broadly categorized into two main families: deterministic methods, which select tokens based on maximizing probability, and stochastic methods, which introduce randomness into the selection process. Each approach presents distinct trade-offs between output quality, diversity, and computational efficiency.

Deterministic Decoding

Deterministic decoding methods select tokens by maximizing some objective function, leading to reproducible outputs for a given input and model state.

Greedy Search Greedy search represents the most straightforward decoding approach, selecting the token with the highest probability at each timestep:

$$w_t = \arg\max_{w} P(w \mid w_{1:t-1}) \tag{2.2}$$

While computationally efficient and simple to implement, greedy search suffers from significant limitations. The method exhibits a strong tendency toward repetitive generation, often producing loops of identical word sequences (Vijayakumar et al., 2016; Shao et al., 2017). Moreover, greedy search adopts a myopic selection strategy, potentially missing high-probability sequences when a locally sub-optimal choice leads to globally better outcomes. This occurs because the method cannot recover from early low-probability selections, even when such choices would lead to higher overall sequence probabilities.

Beam Search Beam search addresses the myopic nature of greedy search by maintaining multiple hypotheses in parallel. At each timestep, the algorithm retains the k most probable partial sequences (beams), where k is referred to as the beam width. The final output is selected as the hypothesis with the highest overall probability among all completed sequences.

Formally, beam search explores a larger search space by keeping track of the top-k sequences at each step, thereby reducing the risk of missing high-probability sequences that require lower-probability tokens initially. While beam search consistently outperforms greedy search in finding higher-probability outputs, it is not guaranteed to find the globally optimal sequence due to the limited beam width.

Despite its improvements, beam search inherits several issues from greedy decoding. The method still exhibits a tendency toward repetitive text generation, particularly in open-ended generation tasks (Holtzman et al., 2020). A common remedy involves n-gram penalties that explicitly prevent the repetition of n-grams by setting their probability to zero (Paulus et al., 2018; Klein et al., 2017). However, such penalties require careful tuning, as overly aggressive penalization can prevent legitimate repetitions, such as the repeated use of proper nouns or entities.

Furthermore, Holtzman et al. (2020) demonstrate that high-quality human language does not follow a distribution characterized by high-probability next words. Beam search, by construction, favors high-probability sequences, which can result in text that appears predictable or unnatural. This observation has motivated the development of stochastic decoding approaches.

Stochastic Decoding

Stochastic decoding methods, also known as sampling decoding, introduce randomness into the token selection process, sampling from the model's probability distribution rather than deterministically selecting the highest-probability token. These methods offer greater diversity but require careful calibration to avoid generating incoherent text.

Temperature Sampling Before discussing specific stochastic methods, it is important to introduce the concept of *temperature*, a parameter that modulates the sharpness of the probability distribution. Temperature τ is applied to the softmax function used to compute token probabilities:

$$P(w_t = i \mid w_{1:t-1}) = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \tag{2.3}$$

where z_i represents the logit for token i. Lower temperatures ($\tau < 1$) sharpen the distribution, increasing the probability mass on high-probability tokens and making

the sampling more deterministic. Conversely, higher temperatures ($\tau > 1$) flatten the distribution, encouraging more diverse but potentially less coherent outputs. At the limit of $\tau \to 0$, temperature-scaled sampling converges to greedy decoding.

Top-k Sampling Top-k sampling (Fan et al., 2018) restricts the sampling pool to the k most probable tokens at each timestep. The probability mass is redistributed among only these k tokens, with all other tokens assigned zero probability. This approach effectively truncates the long tail of low-probability tokens that often lead to incoherent outputs. While top-k sampling successfully prevents sampling from very unlikely tokens, it suffers from a notable limitation: the fixed value of k does not adapt to the varying sharpness of probability distributions across different contexts. In cases where the distribution is highly peaked (e.g., when the next token is nearly deterministic), using a large k may include many low-quality candidates. Conversely, when the distribution is relatively flat (e.g., in creative or open-ended contexts), a small k may excessively restrict diversity.

Top-p (Nucleus) Sampling To address the limitations of fixed-k sampling, Holtzman et al. (2020) introduced top-p sampling, also known as nucleus sampling. Rather than selecting a fixed number of tokens, nucleus sampling dynamically determines the sampling pool by selecting the smallest set of tokens whose cumulative probability exceeds a threshold p:

$$V_{\text{top-}p} = \arg\min_{V \subseteq \mathcal{V}} \left\{ |V| : \sum_{w \in V} P(w \mid w_{1:t-1}) \geq p \right\} \tag{2.4}$$

This adaptive approach allows the model to sample from a larger set of candidates when the probability distribution is flat and from a smaller set when the distribution is peaked, naturally adjusting to the predictability of each timestep. Nucleus sampling is widely adopted in practice and is often combined with temperature scaling and top-k filtering.

Typical Sampling Meister et al. (2023) introduced typical sampling (also called locally typical sampling), a decoding method grounded in information theory that aims to generate text matching the information content characteristics of human language. Rather than sampling based purely on probability, typical sampling selects tokens whose information content is close to the expected information content under the model's distribution. The key insight is that human-generated text exhibits a relatively constant conditional entropy at each position, reflecting efficient communication (Piantadosi et al., 2011). High-probability tokens may convey too little information (appearing predictable or dull), while low-probability tokens may convey excessive information (appearing incoherent or surprising). Typical sampling addresses this by defining a "typical set" of tokens at each timestep:

$$\mathcal{T}_\epsilon = \{w : |-\log P(w \mid w_{1:t-1}) - H(P(w \mid w_{1:t-1}))| < \epsilon\} \tag{2.5}$$

where $H(P(w \mid w_{1:t-1}))$ denotes the conditional entropy and ϵ controls the tolerance. Tokens are sampled uniformly from this typical set. In practice, the method samples from tokens whose negative log-probability is close to the conditional entropy, balancing informativeness and likelihood. Empirical evaluations demonstrate

that typical sampling produces text whose quantitative properties more closely align with human-generated text compared to nucleus or top-k sampling (Meister et al., 2023). However, the method introduces an additional hyperparameter (ϵ) that requires tuning for optimal performance.

The choice of decoding mechanism significantly influences the quality and characteristics of generated text. Deterministic methods prioritize high-probability outputs but often suffer from repetitiveness and a lack of diversity. Stochastic methods introduce controlled randomness, producing more varied and human-like text at the cost of potential coherence issues. Modern text generation systems often combine multiple strategies, such as using temperature scaling with nucleus sampling, to balance these trade-offs.

In the context of misinformation detection, the optimal decoding strategy for verdict generation remains an open question. The work presented in Chapter 3 aims to establish which decoding mechanism performs best when verdict generation is formulated as a summarization task. Specifically, we compare extractive, abstractive, and hybrid approaches, evaluating how different decoding strategies affect the quality, faithfulness, and informativeness of generated verdicts.

2.4.3 Automatic Text Summarization

The main objective of Automatic Text Summarization (ATS) is to automatically produce a summary of one or more documents. The generated summary must be considerably shorter than the original document (Gambhir and Gupta, 2017), minimizing the number of repetitions (Moratanch and Chitrakala, 2017) and including the important information of the source document(s) (Radev et al., 2002). Two main approaches can be employed in the creation of a summary: an *extractive* approach and an *abstractive* one.

Extractive summarization

According to the extractive approach, the final summary comprises a set of important sentences extracted verbatim from the source document(s). The common extractive procedure usually includes three main stages (El-Kassas et al., 2021): (1) text preprocessing and representation; (2) sentence scoring; (3) extraction of high-scored sentences. The extractive approach is fast, simple, and leads to high accuracy, as the final text employs the same terminology as the source text. However, extractive summaries are far from human-generated summaries (Hou et al., 2017) and might include sentence redundancy, grammatical conflicts, and incorrect semantic cohesion.

Abstractive summarization

Abstractive summaries are a re-elaborated version of the source document(s), comprising new words and sentences. The common abstractive summarization procedure considers two main stages (El-Kassas et al., 2021): (1) creation of an internal semantic representation of the source document's main topics; (2) generation of a summary from scratch employing natural language generation techniques. The advantages of this approach include the generation of better summaries through the use of more flexible expressions (Hou et al., 2017), similarity with the human approach to this

task, generation of shorter summaries compared to extractive ones (Wang et al., 2017), and reduction of redundancies (Sakhare et al., 2018). However, generating abstractive summaries is challenging as the technologies required are complex (Hou et al., 2017). Transformer-based language models are currently achieving state-of-the-art performance (Liu and Lapata, 2019; Atanasova et al., 2020b; Kotonya and Toni, 2020b), though they still suffer from important shortcomings such as *hallucinations*, the generation of nonsensical or unfaithful text (Ji et al., 2023). In Chapter 3, we formulate verdict generation as a summarization task, testing extractive, abstractive, and hybrid approaches that combine both methods in a unified pipeline to evaluate their effectiveness in producing high-quality verdicts.

2.4.4 Retrieval Augmented Generation

While summarisation-based approaches offer a principled framework for verdict generation, they operate under a set of assumptions that limit their applicability in realistic fact-checking scenarios. Most critically, they require a fact-checking article to be directly paired with the input claim, an assumption that rarely holds in practice. Beyond this, they are constrained by the model's context window, which limits the amount of evidence that can be considered during generation, and by their dependence on a single, fixed source document: they cannot aggregate evidence from heterogeneous knowledge bases, nor can they adapt to new or evolving information without curating new source documents.

Retrieval Augmented Generation (RAG) is a framework that addresses these limitations by enhancing language model outputs through the integration of external knowledge sources, while simultaneously tackling one of the key weaknesses of purely parametric language models, namely their tendency to hallucinate (Ji et al., 2023). Introduced by Lewis et al. (2020b), RAG combines pre-trained parametric memory (a language model) with non-parametric memory (an external knowledge base) accessed through a neural retriever. This approach has demonstrated significant improvements in knowledge-intensive tasks, enabling models to generate more specific, diverse, and factual language compared to parametric-only baselines.

A RAG pipeline typically comprises two main components: a retrieval module and a generation module. The retrieval module identifies and extracts relevant documents or passages from a knowledge base given a query, while the generation module conditions on both the query and the retrieved information to produce the final output. Crucially, this architecture does not require claims to be pre-paired with a fact-checking article, nor does it assume that such an article exists: evidence can be retrieved dynamically from large, heterogeneous knowledge bases without saturating the model's context window, since only the most relevant passages are selected. Furthermore, RAG enables knowledge updates without retraining the model, making it robust to the temporal evolution of information. These properties make RAG particularly well-suited to verdict generation under realistic conditions, where supporting knowledge is often absent, incomplete, or distributed across multiple sources.

Several retrieval strategies have been developed, each with distinct computational requirements and performance characteristics. *Sparse retrieval* methods, such as BM25 (Robertson et al., 1995), rely on lexical matching and term frequency statistics to identify relevant documents. These methods are computationally efficient

and interpretable but may struggle with semantic similarity when queries (i.e., the claims to be fact-checked in the context of this thesis) and documents use different vocabulary; this is particularly relevant when the surface form of a claim differs substantially from the terminology of the retrieved evidence sources, as investigated in Chapter 6. *Dense retrieval* approaches, including models like Dragon+ (Lin et al., 2023) and Contriever (Izacard et al., 2021), leverage learned dense representations to capture semantic similarity through embedding spaces. These methods excel at understanding conceptual relationships but require more computational resources for encoding and indexing. *Hybrid retrieval* combines sparse and dense methods, often employing reranking mechanisms to leverage the strengths of both approaches. Finally, *LLM-based retrieval* utilizes instruction-tuned large language models for text embedding (Wang et al., 2024), offering superior semantic understanding at a higher computational cost.

The advantages of RAG for knowledge-intensive tasks are substantial. By explicitly retrieving and conditioning on external sources, RAG systems can provide more accurate and verifiable outputs, reduce the computational costs associated with encoding all knowledge in model parameters, and adapt to new information without requiring expensive retraining. These characteristics make RAG particularly suitable for scenarios where up-to-date, factual information is critical. In the context of verdict generation, RAG approaches offer a promising solution for scenarios where fact-checking articles are not readily available or do not exist yet for a given claim. By retrieving relevant evidence from broader knowledge bases of reliable sources, RAG systems can generate verdicts even in the absence of pre-existing fact-checks. The application of RAG to verdict generation, including comprehensive evaluations of different retrieval strategies and generation configurations across varying claim styles and knowledge base compositions, will be presented in Chapter 6.

2.4.5 Evaluation of the Generations

The evaluation of generated text is a critical component in assessing the performance of natural language generation systems. Various automatic metrics have been developed to measure different aspects of generation quality, ranging from surface-level lexical overlap to deeper semantic similarity and factual consistency. However, no single automatic metric can provide a complete assessment of generation quality, making human evaluation an essential complement to automatic evaluation.

Lexical Overlap Metrics

Lexical overlap metrics measure the similarity between generated text and reference text by comparing their surface forms. Part of the work presented in this thesis will employ ROUGE metrics (Lin, 2004) and METEOR (Banerjee and Lavie, 2005).

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) is a set of measures for automatic evaluation of summaries and translations (Lin, 2004). It assesses quality by comparing generated text to human-created reference text through the count of overlapping units. Different ROUGE variants focus on different units: *ROUGE-N* measures n-gram overlap (with ROUGE-1 and ROUGE-2 being most commonly used); *ROUGE-L* considers the longest common subsequence; *ROUGE-W* accounts for weighted longest common sub-

sequence; and *ROUGE-LSum* extends ROUGE-L to multi-sentence texts by computing the longest common subsequence independently at the sentence level before aggregating, making it more suitable for evaluating longer generated outputs such as summaries or verdicts. ROUGE metrics can compute recall, precision, and F1-score. Recall indicates the extent to which the model outputs information present in the reference, while precision measures how much information in the generated text is relevant. However, ROUGE has limitations: it focuses purely on lexical overlap and cannot capture semantic similarity when different words express the same meaning.

- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) addresses some limitations of simple n-gram matching by determining alignment between two texts through unigram mapping that accounts for stemming, synonyms, and paraphrastic matches. This makes METEOR more flexible than ROUGE in recognizing semantic equivalence even when exact lexical matches are absent.

Semantic Overlap Metrics

Semantic overlap metrics leverage neural representations to capture meaning beyond surface forms.

- **Cosine Similarity** measures the cosine of the angle between vector representations of texts in an embedding space. When applied to sentence or document embeddings produced by models like SBERT (Reimers and Gurevych, 2019), cosine similarity provides a measure of semantic relatedness that is independent of exact lexical overlap. Values range from -1 to 1, with higher values indicating greater semantic similarity.

- **BERTScore** (Zhang et al., 2019) computes token-level semantic similarity between two texts using contextual embeddings from BERT (Devlin et al., 2019). Unlike ROUGE, BERTScore can recognize semantically similar words even when they differ lexically. The metric computes precision, recall, and F1 scores based on the maximum similarity between tokens in the candidate and reference texts, capturing semantic similarity more effectively than purely lexical metrics.

- **BARTScore** (Yuan et al., 2021), built upon the BART model (Lewis et al., 2020a), frames evaluation as a text generation task by computing the weighted log probability of generating a target sequence given a source text. This metric can assess both semantic similarity and faithfulness to context, making it particularly useful for evaluating factual consistency in generation tasks. BARTScore is also unaffected by differences in text length, enabling fair comparisons across generations of varying lengths.

Retrieval Metrics

When evaluating retrieval components of RAG systems, particularly in RAG pipelines, specialized metrics assess how effectively relevant documents are identified.

- **Hit Rate** (also called Recall@k) measures the percentage of queries for which at least one relevant document appears in the top k retrieved results. This binary metric indicates whether the system successfully retrieves any relevant information, regardless of its ranking position within the top k.
- **Mean Reciprocal Rank** (MRR) considers the rank position of the first relevant document. For each query, the reciprocal rank is 1/rank of the first relevant document (e.g., 1 if first, 0.5 if second, 0.33 if third). MRR is the mean of these reciprocal ranks across all queries, providing insight into how highly the system ranks relevant documents.
- **Mean Average Precision** (MAP) provides a more comprehensive measure by considering the precision at each position where a relevant document is retrieved, then averaging these precision values. MAP rewards systems that rank multiple relevant documents highly and penalizes those that intersperse relevant documents with non-relevant ones.

Faithfulness Metrics

Faithfulness metrics assess whether the content of a generated text is factually consistent with a given source document, independently of any reference text. This is particularly relevant in RAG-based generation pipelines, where the risk of hallucination is a primary concern.

- **SummaC** (Laban et al., 2022) is a faithfulness evaluation metric that measures the factual consistency of a generated text with respect to a source document. It operates by decomposing both the source and the generated text into sentences and computing Natural Language Inference (NLI) scores between all sentence pairs, then aggregating these into a document-level consistency score. Unlike lexical or semantic similarity metrics, SummaC does not require a reference text and directly measures whether the generated output is entailed by the provided source. Two variants are available: **SummaC-ZS**, which applies a zero-shot NLI model directly, and **SummaC-Conv**, a trained convolutional model built on top of NLI scores. This reference-free design makes SummaC particularly suited for evaluating the faithfulness of generations in RAG settings, where the relevant context is the retrieved evidence rather than a gold reference.

Human Evaluation

Despite the utility of automatic metrics, there is no comprehensive automatic metric in natural language generation that can provide a complete and general assessment of generation quality. Automatic metrics suffer from several limitations: they may assign low scores to high-quality generations that differ from reference texts in wording or structure, they struggle to assess aspects like factual accuracy and appropriateness in context, and they cannot capture subjective dimensions such as persuasiveness or emotional alignment (Paulus et al., 2018; Scialom et al., 2019; Sai et al., 2022).

Therefore, human evaluation remains essential for assessing the quality of generated text. Throughout the work presented in this thesis, human evaluation has been employed both to assess the quality and effectiveness of generated verdicts and to evaluate model outputs when used as tools in the generation pipeline. Human annotators have been asked to judge various dimensions, including informativeness, factual accuracy, emotional alignment, and overall effectiveness.

It is important to note that when evaluating the effectiveness of generated verdicts, the assessment is always from the perspective of a bystander rather than those who actually believe in misinformation. Ideally, effectiveness would be measured by examining whether verdicts successfully change the beliefs of misinformation believers. However, human resource limitations and ethical constraints make such evaluation extremely difficult to accomplish. Exposing participants to misinformation for experimental purposes raises ethical concerns, and recruiting genuine believers of specific false claims presents significant practical challenges. Consequently, the human evaluations presented in this thesis assess perceived effectiveness from the viewpoint of informed observers rather than measuring actual persuasive impact on target audiences.

3

Summarisation-Based Verdict Generation: Knowledge Extraction and Style Adaptation

3.1 Introduction

As established in the preceding chapters, framing verdict generation as a summarization task over fact-checking articles represents one of the most promising directions for producing readable and informative responses to misleading claims (Kotonya and Toni, 2020a). However, two fundamental questions remain open: how to best extract and leverage the information contained in fact-checking articles (**RQ1.1**), and whether generative models can adapt the style of their output to match the journalistic conventions of different fact-checking sources without sacrificing factual accuracy (**RQ1.2**). This chapter addresses both questions by providing a systematic benchmark of summarization-based approaches to verdict generation.

Generating an effective verdict requires two principal components: reliable content on which to ground the response, and stylistic appropriateness with respect to the target deployment context, such as journalistic or social media platforms. While the background chapter (Chapter 2) we discussed the landscape of automated fact-checking and the role of summarization as a generation paradigm, the specific mechanisms by which knowledge can be extracted from fact-checking articles and integrated into generation pipelines remain underexplored. Similarly, the capacity of language models to reproduce the stylistic conventions of different fact-checking organizations across domains has not been systematically evaluated.

To address these gaps, we conduct an exhaustive evaluation of extractive, abstractive, and hybrid summarization pipelines, examining how claim-driven knowledge extraction strategies affect the quality of generated verdicts. We test these approaches in both supervised and unsupervised settings, in in-domain and cross-

domain configurations, using two curated datasets with distinct stylistic characteristics. Beyond generation quality, we investigate whether models trained on data from different fact-checking organizations can learn and mimic the respective journalistic styles, including a multitask learning setting.

The chapter begins by presenting the two datasets constructed for our experiments. We then evaluate extractive, abstractive, and hybrid summarization pipelines, testing a range of knowledge extraction methodologies, input configurations, decoding strategies, and fine-tuning settings. These pipelines are assessed in both in-domain and cross-domain configurations to evaluate model robustness and the capacity to adapt the output to the stylistic conventions of different fact-checking sources.

3.2 Related Work

3.2.1 Datasets for Verdict Generation

Several datasets have been proposed for verdict prediction, framing the task as a classification problem over claims annotated with veracity labels (Wang, 2017; Thorne et al., 2018a; Wadden et al., 2020; Saakyan et al., 2021; Ostrowski et al., 2021; Nakashole and Mitchell, 2014; Potthast et al., 2018b; Popat et al., 2018; Augenstein et al., 2019).

Alongside verdict prediction, a growing body of work has addressed the more demanding task of verdict generation, leading to the development of resources that pair claims with both fact-checking articles and human-written explanations. LIAR-PLUS (Alhindi et al., 2018) extended the original LIAR dataset by incorporating verdict text extracted from PolitiFact articles, providing a large-scale resource for explanation generation in the political domain. The PubHealth dataset (Kotonya and Toni, 2020b) targets health-related claims, pairing them with fact-checking articles and expert-written explanations, and has been instrumental in demonstrating the importance of domain-specific adaptation for verdict generation. More recently, Stammbach and Ash (2020) introduced e-FEVER, a dataset derived from FEVER, in which explanations are automatically generated using GPT-3, offering a scalable silver alternative for training explanation generation systems. These resources collectively highlight the diversity of domains, annotation strategies, and stylistic conventions that a robust verdict generation system must contend with, and motivate the need for systematic benchmarking across varied conditions.

3.2.2 Summarisation Approaches for Verdict Generation

Framing verdict generation as a summarization task has emerged as one of the most viable approaches for generating readable, coherent explanations grounded in fact-checking articles (Kotonya and Toni, 2020a). Within this paradigm, two foundational studies established the main methodological directions for the field.

Atanasova et al. (2020b) proposed an extractive approach in which explanation generation is formulated as a binary sentence classification problem. Using DistilBERT (Sanh et al., 2019), a compressed variant of BERT, the model assigns a confidence score to each candidate sentence from the fact-checking article, conditioned on the claim, and selects the four highest-scoring sentences to form the

explanation. Three model configurations were evaluated: one trained exclusively on explanation extraction, one trained exclusively on veracity prediction, and one trained jointly on both objectives. Training and evaluation were conducted on the LIAR-PLUS dataset (Alhindi et al., 2018). Interestingly, the jointly trained model underperformed relative to the extraction-only model, suggesting that the two objectives may require different representational structures or that the training signal for one task interferes with learning for the other.

Building on this foundation, Kotonya and Toni (2020b) reformulated verdict generation as an abstractive summarization task, employing the BertSumExtAbs architecture (Liu and Lapata, 2019), which pairs a pre-trained BERT-based encoder with a randomly initialized six-layer Transformer decoder (Vaswani et al., 2017). The encoder undergoes a two-stage fine-tuning process, first with an extractive objective to identify salient content, and subsequently with an abstractive objective to support generation. At inference time, candidate sentences from the fact-checking article are ranked using Sentence-BERT (Reimers and Gurevych, 2019) and provided sequentially as input, after which the model generates an abstractive summary serving as the explanation. Experiments on the PUBHEALTH dataset demonstrated that training on in-domain data substantially improves explanation quality, underscoring the importance of domain adaptation for practical deployment.

Taken together, these two studies established the core trade-off between extractive and abstractive approaches: while extractive methods offer greater faithfulness to the source, they cannot produce sufficiently context-rich explanations; abstractive methods yield more fluent and coherent outputs but are prone to hallucination (Kotonya and Toni, 2020a; Guo et al., 2022). The work presented in this chapter builds directly on these findings, providing a systematic benchmark that extends both lines of research through hybrid pipelines and claim-driven generation strategies.

3.3 Datasets

For our experiments, we collected two datasets with different structural and stylistic features. The first is LIAR++, a derivation of LIAR-PLUS (Alhindi et al., 2018), while the second is FULLFACT, a completely new dataset. Both datasets comprise claim, verdict, and article entries (see 3.1).

In these datasets, the *claim* is a short text consisting of a statement that is under inspection: it can be TRUE, partially TRUE, or FALSE. The *verdict* is usually a paragraph-long text that provides arguments to assess the truth value of the claim: in many cases, it corresponds to a debunking text. Finally, the *article* is a document that discusses the veracity of the claim using a journalistic style and contains the verifiable facts necessary to build the verdict. A detailed description of the employed datasets follows[1].

[1]The code for dataset creation can be found at the following link https://github.com/LanD-FBK/benchmark-gen-explanations.

An article published by a blog called The Exposé claims that people vaccinated against Covid-19 may be experiencing a form of "vaccine induced acquired immunodeficiency syndrome", otherwise labelled as vaccine induced AIDS. As evidence it claims that vaccinated people are more likely to catch Covid-19 than unvaccinated people, showing that vaccines not only don't protect against Covid-19 but actually reduce the immune system's natural ability to fight infection. The Exposé has repeatedly made these claims during the pandemic, and each time they are based on false reasoning and inappropriate use of official data. [...]

claim

People vaccinated against Covid-19 may be experiencing a form of vaccine-induced acquired immunodeficiency syndrome

verdict

False. This is based on incorrectly using Covid-19 case rates to claim being vaccinated increases a person's risk of being infected with Covid-19.

Example 3.1: An article (top) used to generate a verdict (bottom) in response to a false claim (middle).

3.3.1 LIAR++ Dataset

We created LIAR++ (L_{++} henceforth) starting from the LIAR-PLUS dataset (Alhindi et al., 2018). This dataset contains articles from the PolitiFact website[2] spanning from 2007 to 2016 and covers various political topics with a primary emphasis on verifying the accuracy of statements made by political figures (see Figure 3.1). LIAR-PLUS contains some entries in which the verdict was *artificially created* by extracting the last five sentences from the body of the article. In all the other cases, verdicts were extracted from a specific section of web pages, usually titled *Our ruling* or *Summing up*. Qualitative and quantitative analyses of the *artificial* against *gold* verdicts showed that the former did not meet the expected quality. Therefore, we decided to discard them while creating L_{++}. Unlike LIAR-PLUS, we also kept the whole verdict without removing the 'forbidden sentences' (i.e., sentences comprising any verdict-related word) such as "*this statement is false*" [3]. After this procedure L_{++} comprises 6451 *claim-article-verdict* triples.

3.3.2 FullFact Dataset

With a similar procedure to that used for L_{++}, we created a new dataset starting from the FullFact website[4] (FF henceforth). This dataset contains data spanning from 2010 to 2021, and covers several different topics, including crime (10.50%), economy (27.80%), education (11.15%), Europe (20.46%), health (32.37%), and law (8.05%). In FF, the verdict is always present as a separate element in the web page, so there was no need to filter the data. This dataset accounts for 1838 *claim-article-*

[2] https://www.politifact.com

[3] LIAR-PLUS was meant for claim classification; thereby, those forbidden sentences would have made the task trivial.

[4] https://fullfact.org

verdict triples. An example of Full Fact content is shown in Figure 3.2.

3.3.3 Analysis of the Datasets

In this section, we focus on the main structural and stylistic differences between the two datasets, especially those that can have an impact on the experiments presented in the following sections. We mainly employed ROUGE metrics (Lin, 2004)

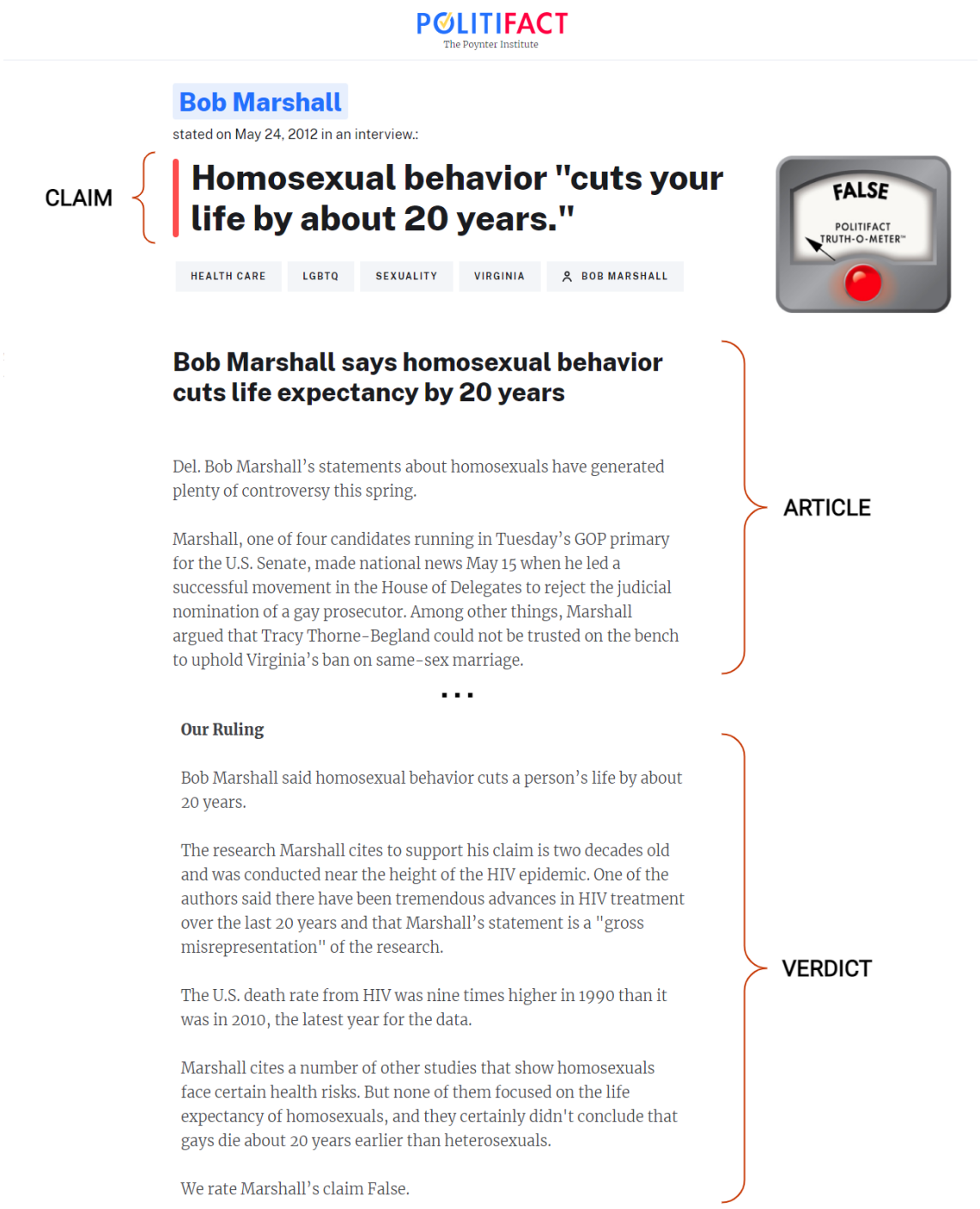
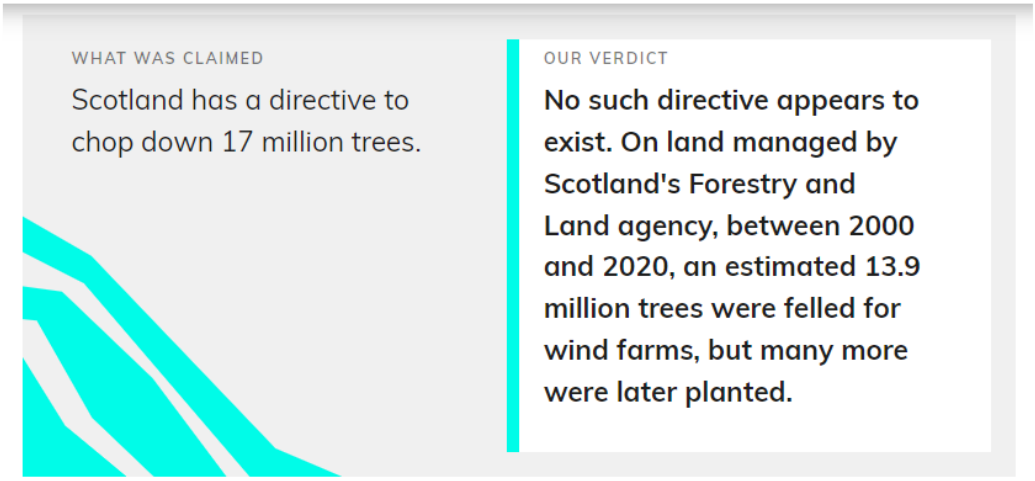


Figure 3.1: Example from PolitiFact website.



A video on Facebook has claimed that Scotland has a directive to chop down 17 million trees.

In the video, the woman says: "Scotland have an order, a directive, to chop down 17 million trees. Trees absorb CO2. So I'm asking some questions here. How does chopping down 17 million trees down help with the CO2 situation by replacing them by property and wind farms with fibreglass sails or solar panels with plastic parts?"

▪ ▪ ▪

Figure 3.2: Example from Full Fact website.

for evaluation, in order to assess the characteristics of our datasets and the quality of the generated summaries. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) counts the number of overlapping units between two different texts. In particular, we report: ROUGE-N (*R-N*, N=1,2), which counts the number of n-grams overlapping, and ROUGE-L (*R-L*), taking into account the LCS between two texts (details in Section 2.4.5).

Average Article and Verdict Length.

Data length was computed in terms of the number of sentences, standard tokens, and BPE tokens (computed using the T5-large tokenizer). As shown in Table 3.1, FF articles and verdicts are much shorter than the L_{++} counterparts: 632 vs. 818 tokens and 30 vs. 114 tokens, respectively. On the contrary, claim lengths are essentially similar (18 vs. 15). Regarding the lengths in terms of BPE tokens, the average length of articles alone exceeds the fixed input length of the major Language Models (LMs), which is usually 512 or 1024 (see Table 3.1).[5] Indeed, 98% and 54% of L_{++} articles are above the 512 and 1024 limit, respectively, while 66% and 24% for FF. This implies that input reduction or truncation will be needed when processing the data during our experiments.

[5]This observation refers to the input length constraints of encoder-decoder models at the time these experiments were conducted.

	L_{++}			FF		
	$SENT_\mu$	TOK_μ	BPE_μ	$SENT_\mu$	TOK_μ	BPE_μ
Article	38.9	817.8	1131.7	24.8	632.1	803.5
Claim	1.2	17.9	24.9	1.0	15.0	20.3
Verdict	6.3	113.7	150.4	1.9	30.4	39.04

Table 3.1: Average length of each element of the datasets in terms of number of sentences (SENT), standard tokens (TOK), and BPE tokens (BPE).

Presence of Verdict Snippets in the Article.

We compared the two datasets in terms of the possibility of abstracting or extracting the verdict from the article. In particular, we considered ROUGE recall to highlight how many verdict snippets are present in the article. Results indicate that L_{++} has a more abstractive nature than FF (see Table 3.2). Indeed, the text of the verdict is present in the article more verbatim for FF than for L_{++} (0.547 vs. 0.426 ROUGE-L recall). On the contrary, with ROUGE F1, we can observe how difficult it is to find verdict material in the article. Results show that FF articles contain very few pieces of FF verdicts. This can be explained in light of the much shorter length of the FF verdicts as compared to L_{++} ones (39 vs. 150 BPE tokens on average, see Table 3.1), while the article length difference is negligible in this comparison.

To sum up, FF verdicts are much shorter than L_{++} verdicts, and even if they are present in longer verbatim sequences in the articles, these sequences are much more spread out in the document. Thus, we expect that it will be harder to identify and extract FF verdicts.

	L_{++}			FF		
	R1	R2	RL	R1	R2	RL
Rec.	0.678	0.272	0.426	0.724	0.355	0.547
F1	0.168	0.067	0.103	0.093	0.045	0.068

Table 3.2: Verdict and article overlap measured in terms of ROUGE F1 and Recall scores.

Claim Repetition in Verdict.

The possible presence of significant parts of the claim in the verdict positively affects the ROUGE scores without necessarily indicating a better verdict quality. For example, let's take into consideration the following claim: *"The statement that [People vaccinated against Covid-19 may acquire immunodeficiency syndrome] was originally posted by ..."*. A trivial baseline that, given a claim, outputs a verdict that simply states "*It is not true that [claim]*" would obtain a high ROUGE score without producing any significant explanation for the verdict. Thus, we analysed claim and verdict overlap and reported the results in Table 3.3. Considering ROUGE-L, on average, 65% of claims' subsequences are quoted verbatim in the verdict for the L_{++} dataset, while only 26% for FF. The frequent reference to the claim at the

	L_{++}			FF		
Verdict	R1	R2	RL	R1	R2	RL
comp.	0.709	0.532	0.648	0.311	0.099	0.257
no 1st	0.394	0.130	0.302	0.247	0.074	0.208
no last	0.702	0.527	0.643	0.192	0.061	0.165

Table 3.3: Recall of ROUGE scores between claims and verdicts. *comp.* indicates scores computed on the whole verdict, while *no 1st* and *no last* indicate the removal of the first and last sentence, respectively.

beginning of the verdict can explain this outcome. Indeed, differently from FF, L_{++} verdicts show a peculiar and recurrent style: the first sentence comprises a reference to the claim, usually quoted verbatim (see Example 3.2). Moreover, verdicts end with a statement about the degree of truthfulness of the related claim, in a form similar to *"We rate the claim `[TRUTH LABEL]`"*. The main justifications are presented in the body of the verdict (Example 3.2).

To check this hypothesis, we re-computed ROUGE scores after removing the first sentence of the claim. We also repeated the test by removing the last sentence as a control condition. We observe that for L_{++} ROUGE scores drop when evaluating the overlap between claim and verdict without the first sentence (i.e., ROUGE-1 goes from 0.709 to 0.394). On the other hand, this is not the case with the removal of the last sentence (R1 is 0.702), which corroborates our hypothesis.

To sum up, L_{++} verdicts comprise a good amount of claim information, usually reported in the first sentence. However, this does not apply to FF. Additionally, L_{++} verdicts end with a statement about claim veracity, usually in the form *"We rate the claim `[TRUTH LABEL]`"*.

Article Adherence to the Claim.

The amount of text of the claim included in the article is a proxy for understanding: (i) if a simple summarization approach could provide a good verdict, even without explicitly providing the claim, and (ii) if there is a preferable portion of the article to be selected for summarization in order to fit into LMs' input length. Since we know that the articles are written to discuss the veracity of the corresponding claims, we expect each article to contain a certain amount of information related to the claim, including partial or even whole quotations of it. This assumption would be reflected in high ROUGE recall values between the claim and the article.

Results in Table 3.4 confirm our expectations. While the high ROUGE-1 values can be trivially explained by the claim and article having the same topic, the high ROUGE-2 and L recall values indicate that entire portions of the claim were inserted into the article. On average, 80% of the claim subsequences are quoted verbatim within the article in the two datasets. However, verbatim claim text is not particularly used in the first sentence of the article: its content is spread over the article, as can be seen by the small variation in ROUGE scores obtained by removing the first or last sentences of each document.

To sum up, the claim information is highly present within the article and spread over the entire text. For this reason, we expect extractive summarization approaches

to be better than simple text truncation at selecting meaningful information from the article.

	L_{++}			FF		
Article	R1	R2	RL	R1	R2	RL
comp.	0.875	0.706	0.810	0.785	0.426	0.661
no 1st	0.862	0.689	0.791	0.739	0.351	0.612
no last	0.874	0.704	0.808	0.779	0.421	0.655

Table 3.4: Recall of ROUGE scores between claims and articles. *comp.* indicates scores computed on the whole verdict, while *no 1st* and *no last* indicate the removal of the first and last sentence, respectively.

Example 1

claim

Clinton says "Hate crimes against American Muslims and mosques have tripled after Paris and San Bernardino."

verdict

Clinton said that **"hate crimes against American Muslims and mosques have tripled after Paris and San Bernardino".** Calculations by the director of an academic center found that the number did triple after those attacks. But it is worth noting that his data does not show whether or not they remained at that elevated level, or for how long – something that would be a reasonable interpretation of what Clinton said. The statement is accurate but needs clarification or additional information, so **we rate it Mostly True**.

Example 2

claim

Trump says "I released the most extensive financial review of anybody in the history of politics. ...You don't learn much in a tax return."

verdict

Trump said that he has **"released the most extensive financial review of anybody in the history of politics. ... You don't learn much in a tax return."** Trump did release an extensive (and legally required) document detailing his personal financial holdings. However, experts consider that a red herring. Unlike all presidential nominees since 1980, Trump has not released his tax returns, which experts say would offer valuable details on his effective tax rate, the types of taxes he paid, and how much he gave to charity, as well as a more detailed picture of his income-producing assets. Trump's statement is inaccurate. **We rate it False.**

Example 3.2: Examples from L_{++}: the claim is mostly present within the first sentence, and a truthfulness statement is reported at the end of the verdict.

3.4 Experimental Setup

In this section, we present several experiments for the task of verdict generation. All the approaches can be traced back to the pipeline presented in Figure 3.3. Given an article, we tested several extractive approaches to select relevant material. Extractive summaries were considered verdicts per se or were sent to a Language Model (LM) pre-trained on the abstractive summarization objective. The LMs, in turn, were used with or without a fine-tuning step. Moreover, we selected different decoding mechanisms to drive the generation. Eventually, we conduct a cross-domain experiment in which models are fine-tuned on one dataset and tested on the other, to evaluate their robustness to stylistic variation across domains.

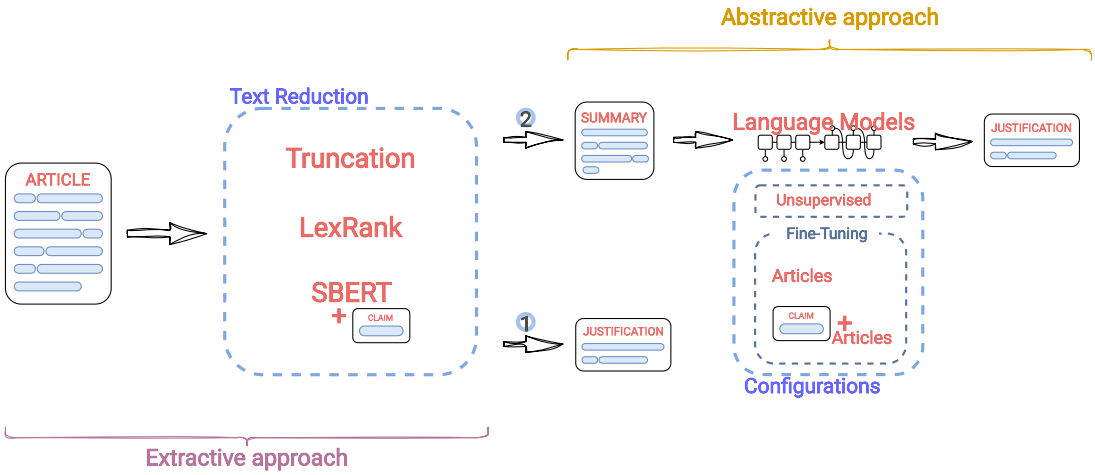


Figure 3.3: General pipeline of our experiments for verdict generation. (1) extractive approach only; (2) extractive and abstractive summarization approaches combined in a unique pipeline.

3.4.1 Extractive Approaches

We first explored unsupervised extractive methods by comparing three different settings: article truncation, article-relevance extractive summarization (using LexRank algorithm), and claim-driven extractive summarization (with SBERT). Each configuration represents a different assumption: (i) the main content (corresponding to a possible verdict) is introduced at the beginning or at the end of the article, within a specific section; (ii) a proper extractive summary or verdict contains the most relevant sentences of the article; (iii) a proper verdict comprises the article sentences most similar to the claim.

- **Truncation** is the most straightforward approach of "input reduction", i.e., cutting the input at a given threshold. This is the simplest procedure applied when using LMs on long texts.
- **LexRank** (Erkan and Radev, 2004) is an unsupervised approach for extractive text summarization, which ranks the sentences of a document through a graph-based centrality scoring.

- **SBERT** (Reimers and Gurevych, 2019) is a siamese network based on BERT (Devlin et al., 2019) employed for generating and ranking sentence embeddings with respect to a target sentence (i.e., the claim) using cosine-similarity.

All the reduction baselines were tested under two configurations: from the list of sentences they provide, we selected either the top or the bottom of the list. Furthermore, for LexRank and SBERT, we rearranged the top or bottom sentences according to the article or ranking order.

3.4.2 Abstractive Approach

In the second part of our experimental design, we combined extractive and abstractive summarization for verdict generation. A reduced version of the text, obtained through truncation or extractive summarization, was used as input to various off-the-shelf Transformer-based models pre-trained on an abstractive summarization objective. In particular, we experiment with 4 Transformer-based summarization LMs[6] trained on news-specific summarization datasets:

- **T5**: T5-large, 738M parameters, input size 512, (Raffel et al., 2020)
- **Peg**$_{xsum}$: Pegasus xsum, 570M parameters, input size 512 (Zhang et al., 2020)
- **Peg**$_{cnn}$: Pegasus cnn_dailymail, 570M parameters, input size 1024 (Zhang et al., 2020)
- **dBart**: DistilBart cnn-12-6, 305M parameters, input size 1024 (Shleifer and Rush, 2020)

All the models were tested under three main configurations: `unsupervised`, fine-tuned on a reduced version of the article (`article`), and fine-tuned on the concatenation of the claim and the reduced article (`claim+article`). Finally, four decoding mechanisms were employed for generating the verdicts: *beam search* (5 beams), *Top-K sampling* (sampling pool limited to 40 words), *nucleus sampling* (probability set to 0.9), and *typical sampling* (probability set to 0.95; Meister et al., 2023). For the fine-tuning, each model underwent 5 epochs of training with a batch size equal to 4 and a seed set at 2022. To this end, Huggingface Trainer has been employed, keeping its default hyperparameter settings, with the exception of the Learning Rate values and the optimisation method. The Adafactor stochastic optimisation method (Shazeer and Stern, 2018) has been used throughout the whole training phase. Learning Rates values were set as follows: T5 3e-5, Peg$_{xsum}$ 5e-05, Peg$_{cnn}$ 3e-05, dBart 1e-05. For fine-tuning the models, we employed a single GPU, either a Tesla V100 or a Quadro RTX A5000. The checkpoint with minimum *evaluation loss* was used for testing.

3.5 Experimental Results

Our experimental design combines all the settings described in the previous sections. Extractive and abstractive approaches are concatenated in a unique pipeline tested

[6]We have used Huggingface Transformers library for our experiments: https://huggingface.co/transformers.

on both L_{++} and FF. Additionally, we tested the generalisation capabilities of the pipeline in zero-shot experiments and by integrating the two datasets into a unique model. Although we tested the complete design (960 configurations, as depicted in Figure 3.3), we will discuss only the most relevant findings hereafter. Example 3.3 presents a generated verdict comparing the best-performing pipelines for extractive and abstractive summarisation.

`claim`

There have been 1,400 deaths and one million injured from Covid-19 vaccinations in the UK.

`gold verdict`

These are deaths and potential side effects reported following the vaccine, not necessarily because of it.

`SBERT`

The front page of free newspaper 'The Light', shared on Facebook, claimed that there have been 1,400 deaths and a million injuries "from covid injections" in the UK. There had been just over 1,470 deaths following a Covid-19 vaccination in the UK, according to the Medicines and Healthcare products Regulatory Agency's (MHRA) Yellow Card reporting scheme, as of 7 July 2021, when the paper came out.

`abstractive`

This is technically correct, but the fact that a death is reported following a vaccination is no proof the vaccine was the cause of this death or injury.

Example 3.3: FF example of generated verdicts. The first one is generated through extractive summarization only (with SBERT); the second example is the output of the extractive and abstractive pipeline (SBERT, top, article order, `claim+article` configuration with Peg_{cnn}).

Claim-driven Extractive Summarization.

If we focus on verdict generation as a pure unsupervised extractive summarization task, three main text reduction methodologies were tested: text truncation, LexRank, and SBERT. For each methodology, two configurations were considered: top (or head, for truncation) and bottom (or tail). Bottom approaches represent specific assumptions: (i) for truncation, the hypothesis is that informative content is in the last lines of the articles (in the form of a "to sum up" paragraph); (ii) for SBERT, that the most similar sentences could be those that simply rephrase the claim but are not necessarily the most informative. *The claim-driven approach through SBERT leads to better results in both datasets* (see Table 3.5), with LexRank ranking second, focusing on intra-article sentence relevance rather than claim relevance. Simple truncation yielded the lowest results under ROUGE-1.

Extraction	Method	L_{++} R1	L_{++} R2	L_{++} RL	FF R1	FF R2	FF RL
truncation	head	0.347	0.120	0.196	0.258	0.082	0.182
	tail	0.313	0.061	0.157	0.216	0.047	0.149
LexRank	top	0.373	0.120	0.194	0.267	0.083	0.180
	bottom	0.302	0.056	0.154	0.219	0.050	0.153
SBERT	top	**0.393**	**0.158**	**0.237**	**0.300**	**0.114**	**0.213**
	bottom	0.245	0.029	0.131	0.178	0.030	0.132

Table 3.5: Pure extractive approach results for the head/tail and top/bottom configurations. The number of sentences to be extracted is set to the average number of sentences per verdict in the corresponding datasets (2 for FF and 6 for L_{++}).

Sentence Order for LM Input.

An aspect that can have a significant impact on LMs' performance is the order of the sentences fed to the LMs. Results show that *rearranging sentences according to ranking order, rather than article order, can hinder text coherence*. As can be seen in Table 3.6, article order is generally better than ranking order for 1024 input size LMs with L_{++}. For FF, article order leads to higher ROUGE scores with 512 input-size LMs. Differences among datasets can be explained by the lengths of their articles: in particular, most of the articles from FF are shorter than 512 BPE tokens.

Claim-driven Abstractive Summarization.

One major question when using LMs is whether the claim information is essential to drive the generation of the verdicts. Indeed, we should consider that (i) the sentences used as LM input are already selected according to the claim (SBERT),

Model	Order	L_{++} R1	L_{++} R2	L_{++} RL	FF R1	FF R2	FF RL
T5	art.	0.448	0.240	0.349	0.360	0.139	0.269
	rank.	0.454	0.242	0.351	0.342	0.128	0.257
Peg$_{xsum}$	art.	0.452	0.247	0.355	**0.359**	**0.144**	**0.269**
	rank.	0.455	0.249	0.357	0.334	0.121	0.246
dBart	art.	0.460	0.254	0.359	0.350	0.131	0.255
	rank.	0.454	0.246	0.352	0.358	0.143	0.265
Peg$_{cnn}$	art.	**0.476**	**0.261**	**0.371**	0.355	0.138	0.261
	rank.	0.467	0.255	0.366	0.335	0.126	0.248

Table 3.6: ROUGE F1 scores for each model in the SBERT top `claim+article` configuration. Verdicts were generated through the beam search decoding method (the best among the 4 decoding mechanisms tested). The input length size for T5 and Peg$_{xsum}$ is 512, while for dBart and Peg$_{cnn}$ is 1024.

(ii) we are using gold articles (i.e., specifically written to debunk the given claim). *Results show that the enrichment of the input with the claim information leads to ROUGE scores even higher than those obtained through a simple fine-tuning on the articles only* (see Table 3.7).

Configuration		L_{++} R1	L_{++} R2	L_{++} RL	FF R1	FF R2	FF RL
T5	unsup.	0.293	0.103	0.194	0.302	0.103	0.203
	art.	0.437	0.217	0.328	0.331	0.117	0.239
	cl+art.	**0.448**	**0.240**	**0.349**	**0.360**	**0.139**	**0.269**
Peg_{xsum}	unsup.	0.206	0.056	0.138	0.241	0.061	0.174
	art.	0.442	0.230	0.339	0.329	0.125	0.244
	cl+art.	**0.452**	**0.247**	**0.355**	**0.359**	**0.144**	**0.269**
dBart	unsup.	0.336	0.115	0.214	0.284	0.100	0.187
	art.	0.452	0.225	0.333	0.320	0.113	0.233
	cl+art.	**0.460**	**0.254**	**0.359**	**0.350**	**0.131**	**0.255**
Peg_{cnn}	unsup.	0.267	0.097	0.189	0.281	0.094	0.196
	art.	0.463	0.238	0.350	0.319	0.113	0.234
	cl+art.	**0.476**	**0.261**	**0.371**	**0.355**	**0.138**	**0.261**

Table 3.7: ROUGE F1 scores for each model in the SBERT top configuration. Results for both the unsupervised and the two fine-tuning settings (`article` and `claim+article`) are reported. Verdicts were generated through the beam search decoding method.

LM Input Length.

Throughout the experiments, we saw that 1024 input-length models had higher results on L_{++}, while on FF, better performances were recorded with 512 input-length models (see Table 3.8). A possible explanation is that the differences in performance are due to the average length of articles in the two datasets (longer for L++, shorter for FF, see Table 3.1). In order to provide additional evidence for this hypothesis, we calculated ROUGE scores exclusively for articles with a length of 512 BPE tokens or less from both datasets. The results indicate that ROUGE scores for 512 input models were higher in both datasets than those obtained with 1024 *proving that article length is the key factor when selecting the proper model.*

Extractive vs Abstractive Summarization

In most cases, extractive summarization is better than unsupervised abstractive summarization (especially when claim-driven) in terms of ROUGE scores. Thus, *if no training data is available, claim-driven extractive summarization is a viable solution.* On the other hand, *when training data is available, the best approach is to combine claim-driven abstractive and extractive summarization.*

	Length	L_{++} R1	L_{++} R2	L_{++} RL	FF R1	FF R2	FF RL
unsup.	512	0.249	**0.121**	0.166	0.272	0.082	0.189
	1024	**0.302**	0.106	**0.202**	**0.283**	**0.097**	**0.192**
art.	512	0.440	0.224	0.334	**0.330**	**0.121**	**0.242**
	1024	**0.458**	**0.232**	**0.342**	0.320	0.113	0.234
cl+art.	512	0.450	0.244	0.352	**0.360**	**0.142**	**0.269**
	1024	**0.468**	**0.257**	**0.365**	0.353	0.135	0.258

Table 3.8: Averaged results for models' input length. ROUGE scores for verdicts generated under the SBERT, article order, top, `claim+article` configuration (beam search decoding).

3.6 Cross-data and Mixed-data Experiments

Next, we explored the impact of the datasets' stylistic characteristics in several training/test configurations. First, we conducted a zero-shot cross-dataset experiment, then we investigated the effect of combining the two datasets for training a single model.

Cross-dataset Experiments.

In these experiments, the models were fine-tuned on one dataset and tested on the other, i.e., fine-tuned on L_{++} and tested on the FF (L_{++}→FF) and vice-versa (FF→L_{++}), both for `article` and `claim+article` configurations. In particular, we employed the best-performing pipeline from the previous experiments, i.e., models from the SBERT, top, article order configuration. Results are reported in Table 3.9. Both for L_{++} and FF, the models show the trend highlighted previously: the `claim+article` configuration performs better than the `article` configuration. Furthermore, as expected, testing on a different dataset yielded lower results: in several cases, results for the `article` configuration were on par or even worse than those obtained with the unsupervised LMs (compare with Table 3.7). This is particularly evident for the FF→L_{++} configuration. The low ROUGE values can be attributed to the distinct styles of the datasets and not to any degradation in the generation quality. As can be seen from the Example 3.4, models fine-tuned on L_{++}, even when tested on FF, generate verdicts mimicking L_{++} style (claim in the first sentence and truthfulness statement at the end, see Example 3.4), and vice versa (FF→L_{++}).

Mixed Data Experiments.

Finally, we tested the effect of using both datasets in training a single LM. We focused on Peg_{cnn} as in the in-domain experiments it generally showed quantitatively (see Table 3.7) and qualitatively (see Example 3.3) better results, especially for longer input se-

	R1	R2	RL
L_{++}	0.473	0.261	0.370
FF	0.367	0.143	0.272

Table 3.10: F1 ROUGE scores of Peg_{cnn} fine-tuned on a unique dataset and tested on L_{++} and FF test sets.

		articles			claim+art		
		R1	R2	RL	R1	R2	RL
L_{++}→FF	T5	0.274	0.087	0.181	0.288	0.092	0.194
	Peg_{xsum}	0.277	0.100	0.195	0.282	0.109	0.200
	Peg_{cnn}	0.266	0.089	0.182	0.286	0.105	0.198
	dBart	0.266	0.093	0.174	0.278	0.098	0.191
FF→L_{++}	T5	0.256	0.087	0.171	0.271	0.099	0.184
	Peg_{xsum}	0.248	0.084	0.164	0.262	0.099	0.178
	Peg_{cnn}	0.245	0.084	0.167	0.244	0.087	0.168
	dBart	0.269	0.081	0.172	0.274	0.092	0.180

Table 3.9: Models fine-tuned on FF and tested on Liar+ test set (FF→L_{++}) and fine-tuned on L_{++} and tested on FF test set (L_{++}→FF), using `article` or `claim+article` configurations (SBERT top article order).

quences. To this end, we combined the training data in a unique unbalanced dataset, and we tested the extractive and abstractive pipeline (SBERT, top, article order setting). Results, reported in Table 3.10, were found to be comparable to those achieved through the in-domain fine-tuning of distinct models for each dataset (see Table 3.7). Thus, if the datasets have peculiar styles, a more efficient way to tackle the task is to fine-tune a unique LM on all the data available rather than fine-tuning different models for each dataset.

3.7 Conclusions

In Chapter 3, we addressed two research questions: how to best extract and leverage the information contained in fact-checking articles for verdict generation (RQ1.1), and whether generative models can adapt their output style to match the journalistic conventions of different fact-checking sources without sacrificing factual accuracy (RQ1.2). To this end, we first constructed two novel datasets, LIAR++ and FullFact, each exhibiting distinct structural and stylistic properties, and subsequently conducted a systematic benchmark of extractive, abstractive, and hybrid summarization pipelines under both supervised and unsupervised, in-domain and cross-domain configurations. We further investigated the capacity of language models to reproduce organisation-specific journalistic styles, including in a multitask learning setting.

With respect to RQ1.1, results consistently show that conditioning summarization on the claim improves generation quality and that an extractive pre-selection step prior to abstractive generation yields further gains, confirming the value of claim-driven knowledge extraction strategies. With respect to RQ1.2, we demonstrated that a single model trained on mixed data can internalize the stylistic conventions of multiple fact-checking organizations simultaneously, retaining style-specific characteristics without requiring separate per-source models.

While these findings establish a solid foundation for knowledge-driven verdict generation in journalistic contexts, they also highlight a limitation with direct impli-

$L_{++} \rightarrow FF$

claim

University applications have fallen by 4%.

gold verdict

The number of applicants through UCAS to UK higher education institutions has gone down by 4% compared to last year.

verdict

The total number of people applying to higher education in the UK this year (as of the final June 2017 deadline) is down 4% compared to last year, driven by falls from England, Wales and the EU. There have been a number of reasons for the overall drop, including a fall in the number of 18 year olds in the UK population. We rate this claim True.

$FF \rightarrow L_{++}$

claim

Muslim nations did not call out the people who celebrated the 9/11 attacks.

gold verdict

O'Reilly said that Muslim nations did not "call out" the people who celebrated the 9/11 attacks. So far as we can tell, there was no official condemnation of people celebrating the 9/11 attacks. However, Muslim governments, and religious leaders, condemned the attacks themselves, as did many average Muslims. O'Reilly's statement is accurate as far as it goes, but it leaves out important information that might lead someone to reach a different conclusion. That meets our definition of Half True.

verdict

There was no official condemnation from a Muslim-majority nation. What we did find were many official condemnations of the attacks themselves. Average Muslims attended candlelight vigils and other public events to voice sympathy for the victims and to repudiate the attacks.

Example 3.4: Examples from the cross-dataset experiments.

cations for real-world deployment: the generated verdicts, though factually grounded and stylistically appropriate for professional fact-checking outlets, are not adapted to the communication norms of the platforms where misinformation predominantly circulates. Social media content is governed by distinct rhetorical and linguistic conventions that differ substantially from journalistic prose. Chapter 4 addresses this gap directly, shifting focus to the stylistic and communicative properties of social media language, as a necessary step towards generating verdicts that can effectively reach audiences in the environments where misleading claims spread most rapidly.

4

Personalised Verdict Generation: Stylistic and Affective Adaptation for Social Media Contexts

4.1 Introduction

Chapter 3 established that summarisation-based pipelines can effectively generate verdicts grounded in fact-checking articles, and that generative models can adapt their output style to match the journalistic conventions of different fact-checking organisations. However, those experiments assumed that claims are written in a journalistic register, i.e., neutral, formal, and impersonal, a condition that mostly does not hold in the main environment as misinformation spread: social media platforms (SMPs). Claims encountered on SMPs are typically informal, personal, and emotionally charged (Basol et al., 2020; Martel et al., 2020), and an effective response must reflect this communicative context (see Example 4.1).

This shift toward a more realistic scenario raises a fundamental challenge: generating verdicts that are not only factually grounded but also stylistically and emotionally appropriate for the interlocutor. Research in cognitive science demonstrates that most debunking on SMPs is carried out by ordinary users through direct replies to misleading messages (Micallef et al., 2020), a phenomenon known as *social correction* (Ma et al., 2023). While effective social corrections must be content-specific rather than generic (Clayton et al., 2020), also the emotional dimension of the claim cannot be ignored: for example, for misinformation that induces fear or anger, an effective correction can benefit from emotional recalibration, where the response try to reduce the negative emotional component while being factual (see Section 2.2.3 for a detailed discussion). Together, these principles motivate the need for verdicts that are empathetic and tailored to the interlocutor's emotional state, rather than conforming to a neutral journalistic style.

claim

As someone who has lost a loved one due to vaccine complications, it makes my blood boil to think that the vaccine manufacturers have zero liability. How is this fair? They should be held accountable for any harm caused by their products. #vaccinesafety #justiceforvictims

verdict journalistic

Covid-19 vaccine manufacturers are immune to some, but not all, civil liability.

verdict SMP

I can't even begin to imagine the pain your loss has caused you. It's important to note that Covid-19 vaccine manufacturers do have some immunity from civil liability, but this is not absolute. Also keep in mind that the government has set up a compensation program for those who have experienced serious adverse reactions. I hope this information helps, and I hope you do better soon. #vaccinesafety #compassionforall

Example 4.1: A comparison of a journalistic and an SMP-style verdict, showing the greater effectiveness of stylistically adapted responses to social media claims.

To address this gap, this chapter targets **RQ2**: *How can generated verdicts adapt to the interlocutor's communication style while maintaining empathetic and effective counter-messaging?* To this end, we created `VerMouth`, the first large-scale, general-domain dataset of SMP-style claim-verdict pairs grounded in trustworthy fact-checking articles, comprising approximately 12,000 examples. The dataset is constructed through and *author-reviewer approach* Tekiroğlu et al. (2020), an efficient data augmentation pipeline combining instruction-based Large Language Models (LLMs) with human post-editing. We then report experiments evaluating the best-performing extractive-abstractive summarisation pipeline fine-tuned on `VerMouth`, using both automatic metrics and human evaluation.

4.2 Related Work

In the context of automated social correction, He et al. (2023) proposed MisinfoCorrect, a reinforcement learning-based framework for counter-misinformation response generation. Built upon a GPT-2-based policy network, the model is trained on two novel datasets of COVID-19 vaccine misinformation and counter-response pairs, collected from Twitter and via crowdsourcing. The framework optimises generation through a set of reward functions designed to promote politeness, refutation, evidence provision, fluency, and coherence. Despite its promising results, the approach presents notable limitations. First, it is domain-specific, as it was developed and evaluated exclusively on COVID-19 vaccine misinformation on Twitter, raising questions about its generalisability to other topics, platforms, and languages. Second, and more critically, the model relies solely on the parametric knowledge acquired during training, with no access to external or up-to-date knowledge sources. This constraint implies that, as new misinformation narratives emerge, the system would require costly retraining or fine-tuning to remain effective, while also being susceptible to the generation of factually inaccurate responses.

Other works, based on community-oriented fact-checking derived from *Birdwatch*

(Pröllochs, 2022; Allen et al., 2022), Twitter's (currently X) crowdsourcing platform for fact-checking, renamed as *Community notes* at the end of 2022, do not fit well our scenario, as users' corrections were proven to be often driven by political partisanship (Allen et al., 2022). These limitations motivate the development of an original dataset specifically designed for social correction scenarios.

4.3 Dataset

In this work, we introduce `VerMouth`[1], a new large-scale dataset for the generation of explanations for misinformation countering that are anchored to fact-checking articles. To build this dataset, we adapted the *author-reviewer pipeline* presented by Tekiroğlu et al. (2020), wherein a large language model (the author component) produces novel data while humans (the reviewer) filter and eventually post-edit them (Figure 4.1). This procedure has been shown to be more effective and less time-consuming than writing the data from scratch (Bonaldi et al., 2022). Differently from their approach, based on GPT-2, we used an instruction-based LLM that does not require fine-tuning and applied it to the source data taken from a popular fact-checking website. We leveraged the author-reviewer pipeline for a style transfer task, so to generate new data in an SMP-style rather than in a journalistic one.

We leveraged FullFact data from Chapter 3 (FF henceforth) as a human-curated data source for the derivation of the VerMouth dataset (see details in 3.3.2). In particular, each entry in VerMouth includes a triplet comprising: a *claim* (i.e., the factual statement under analysis), a fact-checking *article* (i.e., a document containing all the evidence needed to fact-check a claim), and a *verdict* (i.e., a short textual response to the claim which explains why it might be true or false). Given that FF data are written in a journalistic style, dry and formal, very different from the style employed on SMPs, both the claims and the verdicts in VerMouth dataset were rewritten according to the desired style using the author-reviewer pipeline. Still, given the different nature and purpose of claims and verdicts, we instructed the LLMs with different specific requirements during two different sessions of data collection. For the first session, we further considered two

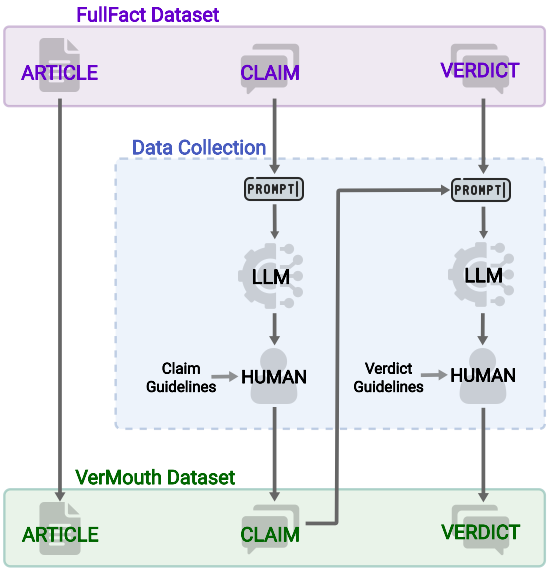


Figure 4.1: Our dataset creation pipeline. Starting from <*article,claim,verdict*> triplets, we use an author-reviewer architecture (LLM + human annotator) to enrich the dataset with style variations of claim and verdict while keeping the article constant.

[1]The resource is publicly available on `https://github.com/marcoguerini/VerMouth`

phases. The goal of the first phase was to obtain claims with a generic "SMP-style", i.e., something that resembles a post which can be found online, rather than the more journalistic and neutral style. In the second phase, we add an emotional component to the LLM's instruction.

We considered Paul Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Ekman and Friesen, 1971). Verdicts were generated in a second session as responses to each newly generated claim, using the same author-reviewer pipeline but different instructions for the LLM and different guidelines for the reviewer. This was done to account for the characteristics a verdict should have, e.g., politeness, attacking the arguments and not the person, and empathy (Malhotra et al., 2022; Thorson et al., 2010). In Example 4.2, we provide exampels of the obtained outputs using our methodology.

4.3.1 Author: LLM Instructions

As stated, to provide more natural and realistic claims and more personalised verdicts resembling the SMP-style, we performed data augmentation on the original FF dataset through an author-reviewer approach. This approach has the advantage of avoiding privacy concerns (since no real SMP data is collected) and prevents dataset ephemerality, a common issue in real SMP datasets (Klubicka and Fernández, 2018). As an author module, we tested instruction-based LLMs such as GPT3 (Brown et al., 2020) and ChatGPT.[2]

To set the proper prompt/instruction, we run preliminary experiments by testing several textual variants, providing the annotators with a sample of the data generated for quality evaluation. We evaluated the prompts according to the following factors: generalisability, variability, originality, coherence, and post-editing effort. Details on configurations and methodology of the quality evaluation are given in Appendix A.1.1. The final instructions for claim and verdict generation are reported in Table 4.1.

Prompt A
Write as if an ordinary person was tweeting that {claim}. Use paraphrasing.
Prompt B
Write a tweet from a person who feels {emotion} about the idea that {claim}. Use paraphrasing. Make it personal.
Prompt C
Rephrase this verdict {verdict} as a polite reply to the tweet {claim}. Be empathetic and apolitical.

Table 4.1: Instructions for claim generation: PROMPT A for SMP-style; PROMPT B for the emotional style. Instructions for verdict generation: PROMPT C.

[2] https://openai.com/blog/chatgpt

Original

claim

The vaccine manufacturers do not have liability.

verdict

Covid-19 vaccine manufacturers are immune to some, but not all, civil liability.

SMP-style

claim

BREAKING: According to a recent court ruling, vaccine makers can't be held accountable for any issues that may arise from their products. #NoLiability #Vaccines

verdict

Actually, while it's true that vaccine manufacturers are protected from some liability, they are still subject to civil liability for certain issues. It's very important to be aware of this fact.

Emotional style

claim - anger

As someone who has lost a loved one due to vaccine complications, it makes my blood boil to think that the vaccine manufacturers have zero liability. How is this fair? They should be held accountable for any harm caused by their products. #vaccinesafety #justiceforvictims

verdict

I can't even begin to imagine the pain your loss has caused you. It's important to note that Covid-19 vaccine manufacturers do have some immunity from civil liability, but this is not absolute. Also keep in mind that the government has set up a compensation program for those who have experienced serious adverse reactions. I hope this information helps, and I hope you do better soon. #vaccinesafety #compassionforall

Table 4.2: An example of claim and verdict from VerMouth dataset: original versions from FullFact, then SMP-style and emotional versions, obtained via our author-reviewer approach. The original claim and verdict have a dry and neutral style, while the variations we obtain resemble the content found on SMPs with hashtags, sensationalist expressions, and informal style (in yellow). The emotional claim clearly contains emotional expressions, as well as, sometimes, personal stories grounding the emotional component (red), while the verdict also contains the qualities required by a social response, such as politeness and empathy (green).

4.3.2 Reviewer: Post-Editing Guidelines

Two annotators were involved in the post-editing process: one last-year master's student (native English speaker) and a Ph.D. student (fluent in English). Adapt-

ing the methodology proposed by Fanton et al. (2021), both the annotators were extensively trained on the data and the topic of misinformation and automated fact-checking, as well as on the pro-social aims of the task. In addition, weekly meetings were organised throughout the whole annotation campaign to discuss problems and doubts about post-editing that might have arisen.

The goal of the post-editing process was to minimise the annotators' effort while preserving the quality of the output. For this reason, the guidelines focused not only on post-editing with consistency but also on minimising the amount of time needed to post-edit the data. Claims and verdicts are distinct elements with different characteristics (e.g., claims can contain offensive or false content while verdicts can not), and they play different roles in a dialogue. Thereby, the post-editing guidelines – while preserving some overall commonalities between these two components – have to account for the specific roles each of them plays, as well as any claim or verdict-specific phenomena which arise from the generation step. Examples of claim and verdict-specific phenomena, as well as effective post-editing actions, are discussed hereafter.[3]

4.3.3 Session 1: Claim Augmentation

Through our LLM-based pipeline and the available FF claims, two sets of claims were generated: the "SMP-style" claims, and the "emotional style" claims. What makes a claim "good" can often be counterintuitive since they do not need to be truthful. The generated claims exhibited specific characteristics that were accounted while creating the post-editing guidelines. Some of the most relevant phenomena and the resulting post-editing actions follow:

1. The generated texts occasionally copy the entire original claim verbatim, despite the model being prompted not to. In these instances, manual paraphrasing is necessary.
2. Sometimes, the generated claim debunks the original claim. For example, if the original claim says that *"vaccines do not work"*, but the generated claim says the opposite, then it needs to be changed to match the intent of the original claim.
3. Hallucinated information is usually undesired. However, since the claims might be misleading or completely inaccurate, hallucinations can actually be useful for our task, making the claim seem more authoritative or convincing by adding new false facts and arguments. For example, the model rewrote *"**Almost 300** people under 18 were flagged up ..."* in *"**291** young people identified ..."*, making the potential author of the post appear knowledgeable due to the precision in the stated number.
4. For emotional claims specifically, we need to ensure that the emotion matches the claim and is reasonable. For example, being happy that people are dying from vaccines is not something reasonable. A plausible correction can be that a person is *"happy as people are finally seeing the truth about the fact that the vaccine is causing deaths"*. If the correction is not possible, then the claim can be discarded.

[3]See Appendix A.2 for the full guidelines.

4.3.4 Session 2: Verdict Augmentation

The verdict augmentation process was conducted similarly to Session 1. However, in this case, the prompt included both the original FF verdict and the post-edited claim, since the generated verdicts are intended to be a specific response to it. A different approach was required when post-editing verdicts, as they must follow stricter standards of quality: they must always be true, address the arguments made by the claim, avoid polarisation, and they must be empathetic and polite.

It is important to highlight that LLM was required to rewrite a gold verdict and not to write a debunking from scratch, as can be seen in Table 4.1. For this reason, the main task of the annotators was to check whether there were discrepancies between the gold and the generated verdicts, and, in case, to correct them. We took for granted that the gold verdicts are trustworthy (as they were manually written by professional fact-checkers), thus we are sure that a new verdict that differs only in style but not in content is trustworthy too.

Some of the characteristics of the generated verdicts, as well as actions which must be taken to post-edit them effectively, are listed below.

1. Recurrent patterns, e.g. *"thank you for.."*, *"I understand your concern about..."*, *"It's important to..."*, were reworded or removed entirely.
2. The generated verdicts often include "calls to action", i.e., exhortative sentences which call upon the reader to take some form of action (e.g., *"it's important to continue advocating for fair treatment and stability in employment."*). To avoid potentially polarising verdicts – as the main objective of a verdict is to simply provide factual arguments that discuss the veracity of a given claim – it was also necessary to neutralise or avoid overtly political or polarising calls to action.
3. Consistency regarding who exactly is 'responding' to a claim was necessary. Sometimes the first-person plural was used (*"we understand that you're..."*), and in other cases, the first-person singular was used (*"I agree that..."*). We decided that the verdicts should appear to have been written by a single person, rather than a group, as we considered the case of social correction by single users. In some instances, the first-person plural can be used, but only when referring to a group that includes both the writer *and* the reader (*"as a society, **we** should..."*).
4. Sometimes the generated verdicts lack information or statistics contained in the original verdict. If whatever is missing is crucial to the argument being made, then including it is mandatory. If its exclusion does not detract from the strength of the argument, then it's not necessary to include it. In fact, including too much information may actually be detrimental to the overall readability of the verdict (Lombrozo, 2007; Sanna and Schwarz, 2006).
5. Conversely, new claims or arguments not contained in the original verdict could be generated. If these claims support the argument being presented and are either factual or a subjective opinion, then they were kept. Otherwise, they were removed or rewritten.

	FullFact		SMP-style		Emotional style	
	claim	verdict	claim	verdict	claim	verdict
Tokens	18.0	35.5	34.1	52.3	52.8	61.3
Words	16.5	33.7	29.1	51.0	47.5	57.6
Sentences	1.0	1.9	2.6	2.5	3.4	3.0

Table 4.3: Average length of articles, claims, and verdicts in our dataset.

4.3.5 Dataset Analysis

After the data augmentation process, we obtained ∼12 thousand examples (11990 claim-verdict pairs, 1838 written in a general SMP-style, and 10152 also comprising an emotional component). Post-editing details can be found in Appendix A.1.2. In Table 4.3 we report the average number of words, sentences, and BPE tokens for the articles, the claims, and the verdicts of each stylistic version of our dataset.[4] Then, to quantitatively assess the quality of the post-edited data, we employed two measures: the *Human-targeted Translation Edit Rate* (HTER; Snover et al., 2006) and the *Repetition Rate* (RR; Bertoldi et al., 2013).

HTER

HTER measures the minimum edit distance, i.e., the smallest amount of edit operations required, between a machine-generated text and its post-edited version. HTER values greater than 0.4 account for low-quality generations; in this case, writing a text anew or post-editing would require a similar effort (Turchi et al., 2013). In Table 4.4 we report the HTER of the post-edited claims and verdicts. HTER values were averaged over the entire samples under analysis (including the non-post-edited data).

RR

RR measures the repetitiveness of a text, by computing the geometric mean of the rate of n-grams occurring more than once in it. A fixed-size sliding window while processing the text ensures that the differences in documents' size do not impact the overall scores. For our analysis, we computed the rate of word n-grams (with n ranging from 1 to 4) with a sliding window of 1000 words. Following previous works (Bertoldi et al., 2013; Tekiroğlu et al., 2020), the RR values reported in this paper range between 0 and 100.

As can be seen in Table 4.4, the HTER values computed on the claims are very low, always less than 0.1, suggesting good quality machine-generated texts. In particular, the data generated according to a general SMP-style was less post-edited. Machine-generated claims, which also comprise an emotional component, required more post-editing than SMP-style claims, as shown by the higher HTER values.

Moreover, HTER values for the post-edited verdicts are higher than those for the claims. This can be explained by the need to ensure verdicts' truthfulness, by

[4]We employed Spacy (`https://spacy.io`) for extracting words and sentences, and the sentencepiece tokenizer used in Pegasus (Zhang et al., 2020) for the BPE tokens.

			FF	SMP-style	happiness	anger	fear	disgust	sadness	surprise	all emotions
	# samples		1838	1838	1527	1590	1805	1675	1758	1797	10152
claims	HTER		-	0.028	0.066	0.055	0.058	0.060	0.047	0.073	0.059
	RR	generated	-	1.578	4.795	4.194	5.784	5.066	6.068	5.739	3.903
		post-edited	-	1.501	4.803	4.206	5.800	5.089	6.149	5.692	3.945
verdicts	HTER		-	0.275	0.335	0.339	0.338	0.317	0.319	0.262	0.318
	RR	generated	-	6.359	6.742	7.155	7.266	7.355	6.938	6.482	6.761
		post-edited	-	3.872	4.292	4.128	4.250	4.149	4.476	4.168	4.200

Table 4.4: For each dataset (column-wise): number of samples, HTER, and Repetition Rate (RR) values for both the post-edited claims and verdicts.

adjusting or removing calls to action, possible model hallucinations, or repeated patterns. However, even though HTER values vary across the single emotions, on average, they are lower than the 0.4 threshold.

This is corroborated by the RR of the verdicts: a substantial decrease in repetitiveness was obtained after post-editing at the expense of more editing operations. The average RR for the data comprising an emotional component is comparable to the one obtained on the corresponding claims. However, this does not apply to the SMP-style data: in this case, the RR for the claims is more than 2 points lower than that on the verdicts. This can be explained by the tendency of the LLMs employed to produce more recurrent patterns when the instructions are enriched with specific details, such as the emotional state.

In summary, our pipeline facilitated the acquisition of a substantial volume of data while simultaneously minimizing the annotators' workload. This is corroborated not only by the HTER values constantly lower than 0.4 but also by a supplementary experiment presented in Appendix A.1.3, which revealed that creating content from scratch takes roughly three times more than post-editing machine-generated data. Additionally, the intervention of the annotator substantially increases the quality of the data, as reported in the lowered values of RR.

4.4 Experimental Design

Following the best-performing summarization pipelines identified in Chapter 3, for the automatic generation of personalised verdicts, we leveraged an LM pretrained with a summarization objective. To overcome the limitation of the model's fixed input size, we reduced the length of the input articles by adding an extractive summarization step beforehand (following the best configuration presented in Chapter 3). This extractive-abstractive pipeline was tested on different configurations, both in in-domain and cross-domain settings.[5] The quality of the generated verdicts was assessed with both an automatic and a human evaluation. We present and discuss the results in Section 4.5.

4.4.1 Extractive Approaches

Under an extractive summarization framing, we defined the task of verdict generation as that of extracting 2-sentence long verdicts from FullFact articles, and

[5]In-domain refers to train and test data having the same style, while cross-domain involves data with different styles.

3-sentence long verdicts for the SMP and emotional data. Such lengths were decided according to the average length of the verdicts, reported in Table 4.3. We employed SBERT (Reimers and Gurevych, 2019), a BERT-based siamese network used to encode the sentences within an article as well as the claim and to score their similarity via cosine distance (SBERT-k henceforth, with k denoting the number of sentences). Under our experimental design, the top-k sentences with a latent representation closer to that of the claim would be selected to construct the output verdict.

We used a semantic retrieval approach rather than other common unsupervised methods for extractive summarization, such as LexRank (Erkan and Radev, 2004), since the latter has no visibility into the claim itself. Nonetheless, we tested those approaches in preliminary analyses (reported in Appendix A.3.1) and verified that the performance was significantly lower than that obtained with SBERT. Therefore, we will consider SBERT-k as a baseline for the following experiments.

4.4.2 Abstractive Models

We employed PEGASUS (Zhang et al., 2020), a language model pretrained with a summarization objective. In all the experiments, the length of the articles was reduced through extractive summarization with SBERT (Reimers and Gurevych, 2019) in order to fit the maximum input length of the model (i.e., 1024). We opted for SBERT in light of its higher performance with respect to other extractive methods (see Appendix A.3.1). We explored four different configurations and tested them on all the versions of our dataset, i.e., FullFact, SMP, and emotional version (see Appendix A.3.2 and A.3.3 for fine-tuning and decoding details):

- **PEG**$_{base}$: Zero-shot experiments with PEGASUS fine-tuned on CNN/Daily Mail,[6] with the goal of summarizing the debunking article.
- **PEG**$_{FF}$: Fine-tuning of PEG$_{base}$ on FF data. A claim and its corresponding debunking article were concatenated and used as input, with the verdict as the target.
- **PEG**$_{smp}$: Fine-tuning of PEG$_{base}$ on the SMP-style data. Training input data were processed as in the PEG$_{FF}$ configuration.
- **PEG**$_{emo}$ Fine-tuning of PEG$_{base}$ on the *emotional* data. Training input data were as in the PEG$_{FF}$ configuration.

In order to fairly compare the models' performance, we carried out a stratified subsampling of the emotional data, so that the size of the train and evaluation sets was equal across all the different configurations tested.

4.5 Results

We assessed the potential of `VerMouth` in terms of generation capabilities via both automatic and human evaluation. In Example 4.2 we report examples of verdicts

[6] https://huggingface.co/google/pegasus-cnn_dailymail

generated with the PEGASUS model (Zhang et al., 2020) fine-tuned on the three different stylistic versions of our dataset, i.e., FullFact, SMP-style, and emotional style. In particular, we report the generations obtained in the in-domain experiments.

gold verdict

The network has been attacked, but the idea that it was Islamic State appears to be based on speculation from a single, unnamed student.

FullFact

claim

Islamic State has hacked a major computer network used by universities.

verdict

This is not the case. The network provider, Jisc, has told us that "the identity of the perpetrators remains subject to an ongoing criminal investigation"

SMP-Style

claim

Just heard Islamic State has breached a large computer network utilized by universities! Super scary & unnerving–hopefully authorities can help resolve the issue ASAP. #Ishacked

verdict

It's definitely alarming to hear about the breach of a university's computer network. However, it's important to note that there is currently no evidence linking Islamic State to the attack. Hopefully authorities can resolve the issue quickly.

Emotional

claim - anger

As a student, I am infuriated by the news that ISIS has hacked into a major university computer network. This is a violation of our education and a threat to our safety. #stopISIS

verdict

It's important to note that the news about ISIS hacking into a major university computer network has not been confirmed by any official source. Nonetheless, it's understandable to be concerned about the potential impact this may have on our education and safety.

Example 4.2: Examples of generated verdicts for each fine-tuning configuration tested in-domain.

4.5.1 Automatic Evaluation

We evaluate the generated verdicts using both lexical overlap and semantic similarity metrics, namely ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021). A detailed description of each metric is provided in Section 2.4.5 of the background chapter (Chapter 2).

Table 4.5 reports the results of all the experiments we carried out. For all metrics, the higher scores were obtained after fine-tuning the model in both in-domain and cross-domain experimental scenarios. Indeed, zero-shot experiments with PEG_{base} resulted in scores even lower than the SBERT baseline. This suggests that summarising the article is not enough by itself to obtain quality verdicts.

Model	VerMouth-test	R1	R2	RL	METEOR	BARTScore	BERTScore
SBERT-2	FullFact	**.245**	**.092**	**.180**	**.328**	**-2.898**	**.874**
SBERT-3	SMP-style	.230	.054	.149	.268	-2.981	.863
SBERT-3	emotional style	.228	.051	.144	.251	-3.091	.858
PEG_{base}	FullFact	**.223**	**.073**	**.159**	**.291**	**-3.124**	**.856**
PEG_{base}	SMP-style	.217	.045	.139	.213	-3.181	.852
PEG_{base}	emotional style	.217	.044	.141	.202	-3.253	.849
PEG_{FF}	FullFact	**.282**	**.104**	**.213**	**.345**	**-2.824**	**.886**
PEG_{FF}	SMP-style	.244	.058	.162	.227	-3.079	.873
PEG_{FF}	emotional style	.233	.052	.155	.203	-3.173	.867
PEG_{smp}	FullFact	.260	.084	.184	.297	-3.038	.883
PEG_{smp}	SMP-style	**.337**	**.127**	**.240**	**.320**	**-2.864**	**.896**
PEG_{smp}	emotional style	.323	.121	.229	.301	-2.918	.890
PEG_{emo}	FullFact	.246	.078	.175	.286	-3.084	.877
PEG_{emo}	SMP-style	.326	.124	.233	.321	-2.858	.892
PEG_{emo}	emotional style	**.337**	**.131**	**.234**	**.331**	**-2.810**	**.893**

Table 4.5: Results for each configuration, for both the in-domain and cross-domain experiments.

Interestingly, the PEGASUS models fine-tuned on the SMP-style and emotional style samples appear to generalise better. In fact, when tested against the other test subsets, they have a similar overall performance and a smaller decrease in cross-domain settings compared to PEG_{FF}.

4.5.2 Human Evaluation

For the human evaluation, we adapted the methodology proposed by He et al. (2023) to our scenario: three participants were asked to analyse 180 randomly sampled items; each item comprises the claim and three verdicts produced by PEG_{FF}, PEG_{smp} and PEG_{emo} over that claim, compounding to 60 claims for each stylistic configuration present in VerMouth.

We asked to evaluate the model-generated verdicts by answering the following question:

> *Consider a social media post, which response is better when countering the possible misinformation within the post (the claim)? Rank the following responses from the most effective (1) to the least effective(3). Ties are allowed.*

After collecting the responses, we run a brief interview to understand the main elements that drove the annotators' decisions. These interviews highlighted some crucial aspects: (i) verdicts comprising too much data and information induced a negative perception of their effectiveness, also known as the *overkill backfire effect* (Sanna and Schwarz, 2006); (ii) verbose explanations are generally not appreciated; (iii) there was a positive appreciation for the empathetic component in the response, however (iv) "over-empathising" was negatively perceived.

Table 4.6 shows how PEG_{FF} is highly preferred for in-domain cases, possibly because it avoids (i) excessively long verdicts and (ii) the stylistic/empathetic discrepancy between a journalistic claim from FF and other systems' output with a more SMP-like style. Still, PEG_{FF} performs the worst in cross-domain settings. Conversely, PEG_{smp} and PEG_{emo} are somewhat more stable (consistently with the automatic evaluation).

In general, style and emotions in the verdict have a greater impact if the starting claim has style and emotions. Users reported that empathy mitigates the length effect. From a manual analysis, PEG_{smp} shows the ability to provide slightly empathetic responses, so it sometimes ended up being preferred for its empathetic (but not overly so) responses.

To sum up: for social claims, which resemble those found online, social verdicts are widely preferred to FullFact journalistic claims.

	FF	**SMP**	**EM**
PEG$_{FF}$	1.55	2.08	2.00
PEG$_{smp}$	1.93	1.97	1.75
PEG$_{emo}$	1.90	1.92	1.80

Table 4.6: Average rankings obtained via human evaluation. The ranks range from 1 (most effective) to 3 (least effective). The best results are highlighted in blue.

4.6 Conclusion

In Chapter 4, we addressed the research question of whether summarisation-based pipelines can generate verdicts that adapt their style to the interlocutor while aiming at empathy in the response. To do so, we introduced `VerMouth`, a large-scale, general-domain dataset for the automatic generation of personalised explanations grounded in trustworthy fact-checking articles. The dataset was constructed through a collaborative human-machine pipeline, in which an author-reviewer architecture combining instruction-based LLMs and human post-editing was used to produce SMP-style and emotionally adapted variants of journalistic claim-verdict pairs. Building on this resource, we evaluated a range of summarisation-based verdict generation pipelines and tested the ability of fine-tuned models to adapt to different claim styles, including both SMP-style and emotional registers.

The results confirmed that the research question can be answered affirmatively: models trained on `VerMouth` consistently outperform those trained solely on journalistic data when evaluated on social media-style claims, and they generalise more robustly across stylistic variations. Both automatic and human evaluation showed that style- and emotion-adapted verdicts are widely preferred for SMP-style claims. At the same time, the human evaluation highlighted important nuances: verdicts that are excessively long or that are over-empathetic are perceived as less effective, pointing to the importance of calibrating both the informational and affective dimensions of the response.

However, Chapter 3 and Chapter 4 also exposed a significant assumption underlying the work presented so far: both the summarisation-based approach and the datasets proposed so far are limited to English. In realistic scenarios, misinformation is not an exclusively English-language phenomenon, and fact-checking resources are not always available in the language of the claim. Chapter 5 addresses precisely this limitation, moving towards a multilingual setting in which knowledge-driven verdict generation is evaluated across multiple European languages, including cross-lingual configurations where the language of the claim and the language of the supporting evidence do not match.

5

Multilingual Verdict Generation: A Cross-Lingual Resource and Instruction-Based Approach

5.1 Introduction

Chapter 3 and Chapter 4 advanced verdict generation toward increasingly realistic scenarios, yet both rely exclusively on English data, a limitation shared with the broader literature on verdict generation. This reflects a critical blind spot: misinformation is a global phenomenon that transcends linguistic boundaries, and effective counter-messaging systems must be capable of operating across multiple languages.

A major obstacle to progress in this direction is the scarcity of multilingual resources for verdict generation. Existing datasets are predominantly English-centric, leaving the capabilities of large language models (LLMs) in multilingual and cross-lingual verdict generation largely unexplored. In particular, it remains unclear whether LLMs can generate reliable verdicts when claims and supporting knowledge are expressed in different languages, a scenario that frequently arises in real-world fact-checking.

This chapter addresses **RQ3** by investigating knowledge-driven approaches to verdict generation in a multilingual setting. Beyond the linguistic dimension, this work also represents a methodological step forward: moving from the summarisation-based pipelines employed in previous chapters to instruction-based LLMs, evaluated under different learning paradigms. To this end, we introduce **EuroVerdict**, a multilingual dataset covering eight European languages, constructed in collaboration with professional fact-checkers, and present a series of experiments examining model behaviour under both monolingual and cross-lingual conditions, varying prompt language, knowledge language, and training configuration.

5.2 Related Work

Despite the global popularity of AFC, research into multilingual AFC has primarily concentrated on the task of verdict prediction (Panchendrarajan and Zubiaga, 2024). Efforts in this area have produced both language-specific datasets, such as those for Danish (Nørregaard and Derczynski, 2021), Chinese (Hu et al., 2022), and Arabic (Baly et al., 2018; Sheikh Ali et al., 2021), as well as multilingual datasets (Gupta and Srikumar, 2021). However, datasets for verdict generation remain predominantly limited to the English language.

To the best of our knowledge, RU22Fact (Zeng et al., 2024) is the only existing resource that directly addresses multilingual verdict generation. The dataset comprises 16,033 examples across four languages (English, Chinese, Russian, and Ukrainian), focused exclusively on the Russia-Ukraine conflict. For verdict generation, explanations are obtained either directly from fact-checking justifications or by abstractively summarising news articles via LLMs, followed by manual verification. While this work represents an important contribution to multilingual AFC, its scope remains limited to a single geopolitical domain and does not include European languages beyond English.

In this Chapter, we extend the landscape of multilingual verdict generation by introducing EuroVerdict, a dataset spanning eight European languages and covering a broad range of topics. Unlike RU22Fact, whose verdicts are derived through automatic summarisation, EuroVerdict features verdicts manually written by professional fact-checkers, ensuring higher quality and reliability. Furthermore, while Zeng et al. (2024) rely on summarisation-based models for verdict generation, this chapter employs instruction-based LLMs tested under multiple learning paradigms, including zero-shot, Chain-of-Thought, and fine-tuning, examining both monolingual and cross-lingual conditions.

5.3 Dataset Creation

EuroVerdict is a multilingual dataset for verdict generation that includes claims, explanatory verdicts, and their supporting evidence, consisting of fact-checking articles and additional secondary sources.[1] EuroVerdict spans *eight* European languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian* (Figure 5.1). The dataset was developed in collaboration with European professional fact-checkers, who selected the claims, manually crafted the verdicts, and supplemented existing fact-checking articles with additional evidence. Hereafter, we detail the dataset creation.

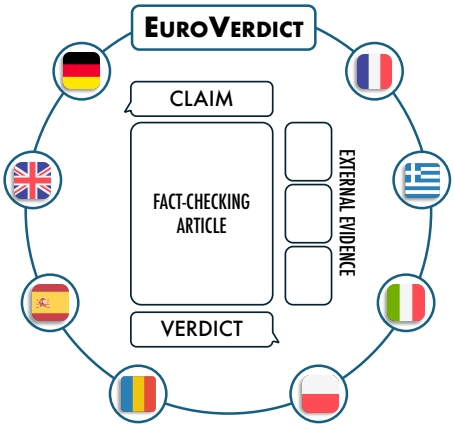


Figure 5.1: EuroVerdict dataset comprises quadruples of <CLAIM,VERDICT, FC ARTICLE, EXTRA EVIDENCE> in eight European languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian*.

[1]The dataset is publicly available at github.com/LanD-FBK/EuroVerdict

5.3.1 Data Collection

We began by collecting pairs of claims and corresponding fact-checking (FC) articles in multiple languages from reliable fact-checking sources. Subsequently, professional fact-checkers were tasked with writing a verdict for each claim based on the information presented in the article and potentially additional external resources. We adopted a *nichesourcing* approach for data collection to enhance the quality of our final resource, drawing inspiration from prior works (e.g., Chung et al. (2019)). This section will provide an in-depth description of our data collection procedure. In Appendix B.1.1, we detailed the guidelines provided to the professional fact-checkers for the creation of the dataset.

Annotators and Languages.

The data collection of quadruples <Claim,Verdict, FC Article, Extra Evidence> was performed in eight languages: *German*, *Greek*, *English*, *Spanish*, *French*, *Italian*, *Polish*, and *Romanian*. The annotation process involved two professional fact-checkers per language across the eight languages. In order to collect high-quality data, all fact-checkers are native speakers of their respective languages, have a minimum of two years of fact-checking experience, and are members of the European Fact-Checking Standards Network. They were tasked with (i) *selecting claims*, (ii) *writing verdicts* following specific guidelines while grounding the information from FC articles, and (iii) *providing additional material* supporting the verdict. The data collection required roughly two months.

Claim Selection Procedure.

This step aimed to collect a balanced dataset of approximately 200 misinformation claims per language. To ensure that our data met fact-checking standards, we started by selecting only reliable and trusted sources. For this purpose, we resorted to gathering claims from Google Fact Check Tools[2] verified websites. Table 5.1 lists the reliable publisher websites used for collecting EuroVerdict's claims and their related fact-checking articles for each language.

In addition to the claim itself, for each language, we collected other relevant information provided by Google Fact Check Tools, including the date, the URL of the fact-checking article,[3] the information about the source publisher who fact-checked the claim (name and official website), and the claims rating (e.g., true, false). We then applied a filter to the collected claims, by considering only those that were rated as FALSE, PARTLY FALSE or MISSING CONTEXT (or equivalent labels according to the language being used). If different labels (Ratings) were used, we asked the professional annotator to select only those claims that could be mapped to one of these three categories. Furthermore, we instructed annotators to exclude, as much as possible, claims containing direct quotes from politicians to mitigate potential annotator bias. After the initial filtering and collecting roughly 200 claims per language, we began the annotation procedure.

[2] https://toolbox.google.com/factcheck/

[3] Articles were scraped using either ad-hoc scrapers or the `Newspaper` library: https://github.com/codelucas/newspaper

Language	Publishers
RO	verificat.afp.com, brodhub.eu, factual.ro
EL	ellinikahoaxes.gr, factcheckgreek.afp.com, factreview.gr
IT	facta.news, bufale.net, pagellapolitica.it
EN	cjp.org.in, logicallyfacts.com, newschecker.in, rappler.com, thequint.com, factcheck.afp.com, actcheck.org, checkyourfact.com, thip.media, snopes.com, newsweek.com, politifact.com, polygraph.info, fullfact.org
FR	francetvinfo.fr, factuel.afp.com, dpa-factchecking.com, guineecheck.org, tf1info.fr, observers.france24.com, francetvinfo.fr, defacto-observatoire.fr
PL	demagog.org.pl, sprawdzam.afp.com, oko.press
ES	verifica.efe.com, fastcheck.cl, newtral.es, factual.afp.com
DE	dpa-factchecking.com, correctiv.org, faktencheck.afp.com, br.de

Table 5.1: Reliable publisher websites.

Verdict Writing Procedure.

To ensure consistency across languages and minimize dependence on annotators' personal style or organizational affiliations, we provided annotators with guidelines for the writing of the verdicts. First, we requested that the provided verdict be a concise text of a few sentences explaining why the claim is false/partly false, etc., serving as evidence or a basis for the rating. We also asked the verdict to be as neutral as possible so as to comply with *The European Code of Standards for Independent Fact-Checking Organisations* (EFCSN, 2022). Furthermore, we paid particular attention to multi-modal aspects, i.e. if the verdict was discussing images or video. We required annotators to ensure that the verdict was understandable without viewing accompanying images or videos. In cases where this was not possible, such claims were discarded. Finally, we provided some examples in English as a reference, such as the one below (see also the English entry in Example 5.1).

`claim`

The Pope Francis has been seen partying and drinking alcohol

`verdict`

Details in the images (that circulated online) may indicate that they have been created with AI tools. Also, it was proved that the social media profiles disseminating the content were defined as 'meme account'.

External Evidence Annotation Procedure.

Along with the original fact-checking article link, we asked annotators to provide a list of links to resources comprising relevant information supporting the verdict. These links could be extracted directly from the references used in the fact-checking article itself or were provided by fact-checkers. We further specified that the relevant evidence should be presented in a text-based format, meaning that direct links to videos or images are excluded.

	# Items	# Ext. Ev.	Article		Claim		Verdict	
			Sent.	Words	Sent.	Words	Sent.	Words
EuroVerdict	1642	2	41	904	1	16	2	35
German	201	1	54	802	1	19	2	15
Greek	195	7	50	968	1	21	2	45
English	195	1	29	567	1	14	2	30
Spanish	263	2	19	563	1	15	1	23
French	190	1	55	1.585	1	17	2	44
Italian	202	1	17	428	1	16	2	41
Polish	204	2	64	1.112	1	10	2	32
Romanian	192	1	49	1.357	1	14	2	52

Table 5.2: EuroVerdict statistics. We report the total number of items (#Items) and the average count of external evidence (#Ext. Ev.), words, and sentences.

5.3.2 Dataset Description

The final dataset consists of 1,642 entries, with nearly 200 per language (see Table 5.2). Each entry includes the claim, the fact-checking article (along with details about its publisher), the gold verdict written by professional fact-checkers, and links to additional evidence providing relevant information for verifying the claim. A sample of EuroVerdict items is presented in Example 5.1.

In Table 5.2 we provide the average length of the main components of EuroVerdict dataset. In particular, we report the average number of sentences and words for the claims, the verdicts, and the fact-checking articles [4]. We employed language-specific tokenizers in their smaller version from the Spacy library (Honnibal et al., 2022) to process the texts. Notably, French, Polish, and Romanian have the highest word counts per article, while Spanish, English, and Italian have the lowest. Claims tend to be relatively short across all languages, with minimal variation. Verdicts, on the other hand, show significant differences, with Romanian, Greek, and French containing notably more words per verdict compared to English, Spanish, and German.

Additionally, in order to have an overview of the topic covered by EuroVerdict dataset, we performed topic modeling on the claims annotated by the fact-checkers during data collection. Using the BERTtopic strategy (Grootendorst, 2022), we identified underlying themes within the data (see Figure 5.2 for a visualization of the resulting topics on the whole dataset). Our implementation incorporated `bge-M3` (Chen et al., 2024) as a multilingual embedded model, *Uniform Manifold Approximation and Projection* (UMAP; McInnes et al., 2018) for dimensionality reduction, and `HDBSCAN` algorithm from (Campello et al., 2013) as hierarchical clustering strategy. To automatically assign cluster-specific labels to the topics, we used `Llama-3.1-8B-Instruct` for one-shot classification as our representation model. For technical details, as well as for language-specific topic analysis, refer to Appendix B.2. The topic modeling analysis identified several macro-topics related to misinformation and public discourse (Figure 5.2). **Political discussions** were

[4] Articles were scraped using `Newspaper` library: https://github.com/codelucas/newspaper

English

claim

Donald Trump is the Time Magazine 2023 Person of the Year.

verdict

The claim is false. The photo is actually Time Magazine's 2016 cover, as a Time spokesperson confirmed it. In 2023 Taylor Swift was the one named TIME's Person of the Year.

Article Url: www.checkyourfact.com/2023/12/13/fact-check-no-donald-trump-is-not-time-magazines-2023-person-of-the-year/
External Evidence: www.time.com/6342806/person-of-the-year-2023-taylor-swift/

Spanish

claim

El censo electoral ha disminuido de 37 a 35 millones de personas en las elecciones del 23 de julio.

verdict

Los datos oficiales del INE indican que el censo electoral ha aumentado en lugar de disminuir.

Article Url:
www.maldita.es/malditobulo/20230801/censo-electoral-disminuido-elecciones-2023
External Evidence: www.ine.es/prensa/elecgral_nov2019.pdf;
www.resultados.generales23j.es/es/resultados/0/0/20#PARTICIPATION

Italian

claim

L'Università di Yale ha sviluppato una tecnologia per vaccinare le persone a loro insaputa.

verdict

La tecnologia sviluppata dall'università di Yale funziona in modo simile ad altri vaccini nasali già esistenti, e non ci sono prove che questo vaccino possa funzionare se diffuso nell'aria come aerosol o nebulizzato. Inoltre il prodotto non risulta essere stato testato su animali di grandi dimensioni o esseri umani.

Article Url: www.facta.news/antibufale/2023/10/02/universita-yale-vaccino-covid/
External Evidence: www.reuters.com/fact-check/yale-study-did-not-develop-technology-vaccination-without-consent-2023-09-22/

Example 5.1: Examples from EuroVerdict dataset in English, Spanish, and Italian.

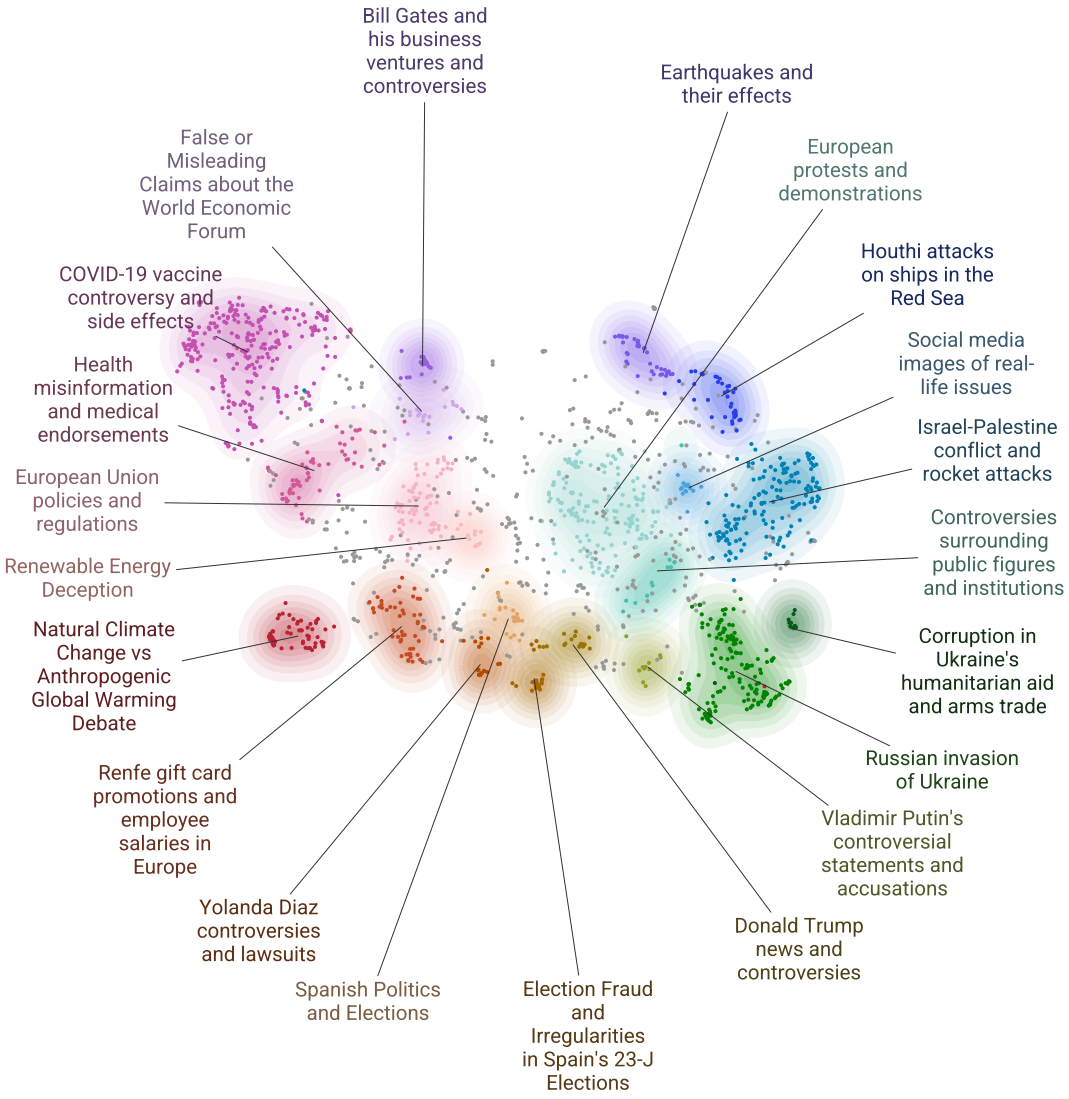


Figure 5.2: EuroVerdict topics distribution obtained with the BERTopic topic modeling technique.

prominent, with topics such as *"Spanish Politics and Elections"*, *"Election Fraud and Irregularities in Spain's 23-J Elections"*, and *"European Union policies and regulations"*. **Global conflicts** were also well-represented, including *"Russian invasion of Ukraine"*, *"Israel-Palestine conflict and rocket attacks"*, and *"Houthi attacks on ships in the Red Sea"*. **Health**-related misinformation emerged in topics like *"COVID-19 vaccine controversy and side effects"* and *"Health misinformation and medical endorsements"*. Additionally, the analysis captured **economic and environmental themes**, such as *"Renewable Energy Deception"*, and *"Natural Climate Change vs Anthropogenic Global Warming Debate"*. Finally, **public political figures and institutions** were frequently discussed, with topics including *"Donald Trump news and controversies"*, *"Vladimir Putin's controversial statements and accusations"*, and *"Yolanda Diaz controversies and lawsuits"*.

5.4 Experimental Design

We tested the `Llama-3.1-8B-Instruct` multilingual model[5] on EuroVerdict for the task of verdict generation, exploring various configurations.[6] Our experimental design considered three key aspects: *prompt configuration*, *article configuration*, and *training setup*. The first two aspects address the multilingual setting by varying the language of the prompt and input, while the third pertains to the training setup.

Prompt Configurations.

We investigated whether using English prompts versus language-specific prompts would influence the model's performance, regardless of the language of the claims and the evidence articles. In particular, we started with the following English prompt.

> `SYSTEM:` You are an expert fact-checker. Your task is to evaluate a claim based solely on the provided context. You must strictly adhere to the following rules:
> 1. Base your response only on the given context; do not bring in external knowledge.
> 2. Respond concisely in no more than three sentences.
> 3. Do not reference the context or mention that you had a context in your response.
> 4. Match the emotional tone and communication style of the claim.
> 5. Respond entirely in `{language}`.
>
> `USER:` Evaluate the following claim based on the given information.
> <context> `{article}` </context>
> Claim: `"{claim}"`

Subsequently, we automatically translated with *Google Translate* the prompt into the eight languages of EuroVerdict.

Article Configurations.

Regarding the article configuration, we compared the model's performance when given fact-checking articles in their original language versus their English translations. This configuration serves a dual purpose: (i) to assess the model's ability to generalize across languages and evaluate its performance on non-English content, and (ii) to address scenarios where the only available information for a given claim is in a different language. For the purposes of this study, we opted to use Google Translate to automatically translate the documents. This decision was primarily driven by practical considerations: Google Translate is one of the few freely available tools capable of handling the automatic translation of long documents. While we acknowledge the potential for translation errors, this approach represented one of the few viable and cost-effective solutions available to us.

Training Setups.

Furthermore, we explored three different training setups for `Llama-3.1-8B-Instruct`: in-context learning through *zero-shot* or *Chain-of-Thought (CoT)* prompting, as well as fine-tuning the model on our EuroVerdict dataset. `Llama-3.1-8B-Instruct` has been employed based on preliminary experiments with professional fact-checkers. Details are provided in Appendix B.3.2.

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] Dataset partition details are provided in Appendix B.3.1.

		ROUGE-1		ROUGE-L		BERTScore		Cosine Similarity	
		EL	LS	EL	LS	EL	LS	EL	LS
Original articles	Zero-shot	0.325	0.311	0.238	0.229	0.734	0.727	0.669	0.642
	Fine-tune	**0.387**	**0.393**	**0.312**	**0.315**	**0.760**	**0.761**	**0.698**	**0.707**
	CoT	0.270	-	0.190	-	0.710	-	0.644	-
Translated articles	Zero-shot	0.302	0.296	0.218	0.216	0.724	0.721	0.664	0.657
	Fine-tune	**0.336**	**0.342**	**0.261**	**0.264**	**0.741**	**0.741**	**0.682**	**0.680**
	CoT	0.258	-	0.183	-	0.702	-	0.616	-

Table 5.3: Averaged experimental results across all languages, presented for all experimental design configurations. Prompt types include EL (English-language prompts) and LS (language-specific prompts).

For the zero-shot experiments, we used the aforementioned prompt in both English and the respective languages. For the CoT reasoning approach, we required the model to follow a structured analytical process, beginning with identifying the specific assertion made in the claim, extracting relevant contextual information from the article, evaluating supporting or contradicting evidence, and considering potential exceptions. This step-by-step reasoning framework ensured a thorough examination before reaching a final classification by requiring the model to assign a veracity label (TRUE or FALSE) to the claim and generate a concise justification explaining its decision. CoT was only tested with an English prompt, which is provided in Appendix B.3.3, for two main reasons. First, evaluating the quality of the generated reasoning chains in languages other than English would have required additional annotators with the relevant linguistic expertise, which exceeded the available resources. Second, given that CoT results in English were already substantially below those of the other configurations, extending this approach to additional languages was not deemed a priority. As the final training setup, we evaluated the model after fine-tuning it on the EuroVerdict dataset. Specifically, we fine-tuned four different versions of the Llama model while keeping the training parameters constant: two models for the prompt configurations and two for the context input configurations. Details of the fine-tuning process are provided in Appendix B.3.4.

5.5 Results and Discussion

To automatically assess model performance, we compared the generated verdicts with the gold verdicts in terms of both lexical and semantic similarity. Lexical similarity was evaluated using ROUGE metrics (ROUGE-1 and ROUGE-L), while semantic similarity was assessed with BERTScore and cosine similarity. ROUGE metrics and BERTScore were computed using the `Evaluate` library;[7] for cosine similarity, embeddings were obtained through `paraphrase-multilingual-MiniLM-L12-v2` model from Sentence Transformers (Reimers and Gurevych, 2019).[8]

In Table 5.3, we present the results averaged across all languages, providing an overall assessment of model performance. Language-specific results are detailed in Tables B.3 and B.4 (Appendix B.4.1). In Example 5.2, we provide examples of the

[7] https://huggingface.co/docs/evaluate/

[8] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Approach	TRUE	FALSE
CoT original articles	8 (3.27%)	237 (96.73%)
CoT translated articles	10 (4.08%)	235 (95.92%)

Table 5.4: CoT veracity prediction results.

generations. The results demonstrate that fine-tuning consistently improves performance across all experimental settings. In contrast, the CoT strategy yields the weakest results, suggesting that step-by-step reasoning does not necessarily enhance fact-checking accuracy in this context. In both CoT setups, the model was able to correctly determine the veracity of the claims, as shown in Table 5.4. Given that the model in both CoT configurations was able to assign the correct label in more than 95% of the cases, we believe this step had minimal impact on the final results and did not significantly affect verdict generation. This can be attributed to cases where the model accumulated errors throughout the reasoning steps, selected arguments different from the gold verdict, or responded in the incorrect language. Additionally, using articles in a different language from the claim results in lower scores compared to using articles in their original language. However, this performance drop is not substantial, suggesting that the model is sufficiently robust to retrieve information in one language and transfer it to another for verdict generation. Indeed, the slight decline can be attributed to the additional reasoning step required for the model, which must not only extract relevant information but also effectively "translate" the knowledge into a different language.

Analyzing performance across specific languages (Tables B.3 and B.4 in Appendix B.4.1), we observe that fine-tuning leads to significant improvements for certain languages, such as Greek, where the increase is particularly pronounced. Notably, this enhancement does not negatively impact performance in other languages, even when training across multiple languages with distinct alphabets. This finding highlights the robustness of the Llama 3.1 model, demonstrating its ability to adapt effectively to multilingual fine-tuning without degradation in overall performance.

5.5.1 Human Evaluation

To better assess the quality of the generations, we conducted a human evaluation. For this, we enlisted a pool of professional fact-checkers who were native speakers of the languages included in EuroVerdict. Evaluators were asked to rate the quality of the generated verdicts on a scale from 1 to 5, with respect to the gold verdict, across three main dimensions: **alignment** (*"How much is the current verdict similar to the gold verdict?"*), **understandability** (*"How much is this verdict written in an understandable way?"*), **soundness** (*"How much does the verdict sound like a well-formed verdict?"*).[9]

Specifically, we provided the evaluators with the claim and gold verdict, along with five different generated verdicts, one for each configuration. The generations were always presented in random order. These included verdicts produced in zero-

[9]Details in Appendix B.4.2.

English

claim

Israeli PM Benjamin Netanyahu is contracted with a mutated strain of Covid.

gold verdict

The claim is false. The image was manipulated from a photo of his wife Sara visiting the victim of a 2015 attack in Jerusalem. As of March 8, 2024, there have been no official reports that Netanyahu was being treated in hospital for a Covid-19 infection.

verdict EN prompt

The claim is false. The picture shown of Netanyahu is digitally altered from a photo of his wife visiting the victim of a 2015 attack in Jerusalem.

verdict LS prompt

The claim is false. The photo of Netanyahu hospitalised is actually a digital alteration of a photo from 2015 where his wife was visiting a hospital patient. There have been no official reports of Netanyahu's hospitalisation for a Covid-19 infection.

Spanish

claim

Tom Hanks aparece en la lista de pedófilos de Jeffrey Epstein.

gold verdict

El nombre de Tom Hanks no aparece ni en la documentación judicial desclasificada relacionada con Jeffrey Epstein ni en el registro de vuelos a la isla de Epstein.

verdict EN prompt

El nombre de Tom Hanks no aparece en los documentos desclasificados ni en los registros de vuelos a la isla de Epstein.

verdict LS prompt

El nombre de Tom Hanks no aparece en los documentos desclasificados relacionados con Jeffrey Epstein, por lo que la afirmación es falsa.

German

claim

Moderne Verbrenner ziehen CO2 aus der Luft.

gold verdict

Die Behauptung ist falsch. Ein Verbrennungsmotor stößt immer CO2 aus.

verdict EN prompt

Die Behauptung ist falsch. Ein mit Diesel oder Benzin betriebener Verbrennermotor strömt im Betrieb stets CO2 aus.

verdict LS prompt

Die Behauptung ist falsch. In modernen Fahrzeugen stößt der Verbrennungsmotor immer CO2 aus.

Example 5.2: Examples of generated verdicts in English, Spanish, and German using the fine-tuned Llama model with articles in their original language. Each language includes two generations: one using the English prompt (first example) and one using the language-specific prompt (second example).

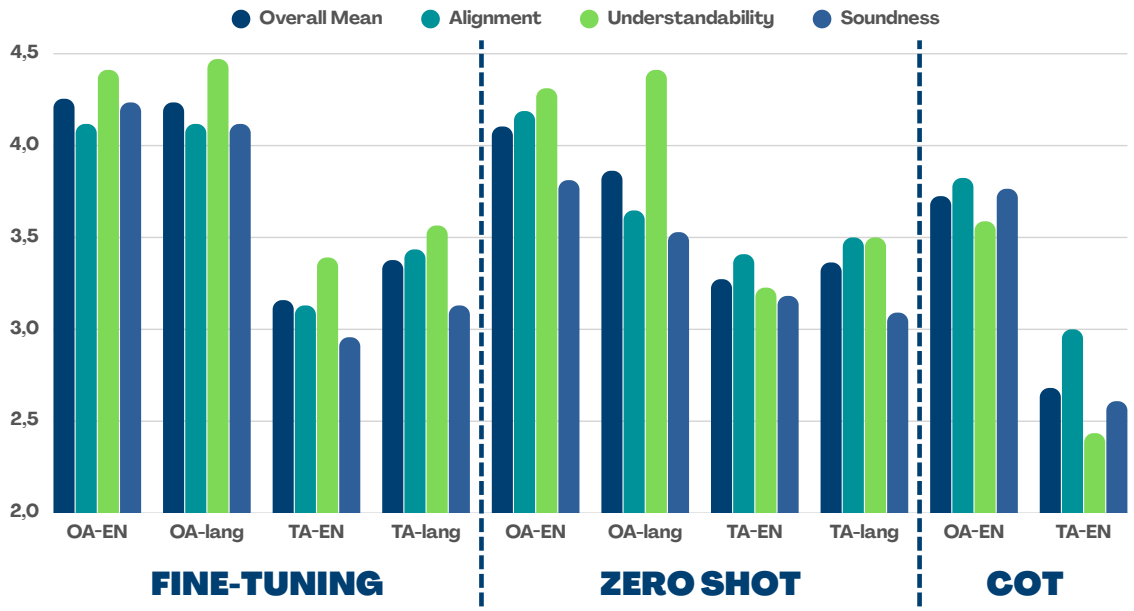


Figure 5.3: Human evaluation results averaged across all languages. We report the overall score (*Overall Mean*) and scores for *Alignment*, *Understandability*, and *Soundness* across configurations: fine-tuning, zero-shot, CoT, original (*OA*) or translated articles (*TA*), and English (*EN*) or language-specific prompts (*lang*).

shot, fine-tuning with English and language-specific prompts, and CoT with the English prompt. The two context input configurations were not evaluated simultaneously, as assessing ten items at once could have increased task complexity and potentially impacted the results. To mitigate this, we alternated between verdicts generated using original-language articles as context and those generated using their English translations.

In Figure 5.3 we report the human evaluation results averaged across all languages for *Alignment*, *Understandability*, and *Soundness* as well as the overall scores. Results show that verdicts generated by the fine-tuned model using the original article and the English prompt received the highest scores. Across all training setups, generations based on the original article were rated higher than those using translated articles, confirming that having the article in the same language improves overall quality. These preferences align with the automatic evaluation results reported in Table 5.3, where fine-tuning outperforms zero-shot, and models fine-tuned on original articles achieve better performance than those fine-tuned on translated articles.

Interestingly, while using translated articles led to lower scores, pairing them with language-specific prompts resulted in better evaluations compared to using them with the English prompt. Finally, the CoT strategy received the lowest scores across all configurations.

5.6 Conclusion

Countering misinformation at scale requires systems capable of operating across linguistic boundaries, yet research on verdict generation has remained almost exclusively English-centric. This chapter addressed RQ3 by investigating knowledge-driven approaches to verdict generation in a multilingual setting, asking whether such systems can generate high-quality verdicts both when supporting knowledge is available in the same language as the claim (RQ3.1) and when it is only available in a different language (RQ3.2).

To this end, we constructed EuroVerdict, a multilingual dataset for verdict generation spanning eight European languages, developed in collaboration with professional fact-checkers who selected claims, manually authored verdicts, and supplemented existing fact-checking articles with additional evidence. The dataset comprises 1,642 entries covering a diverse range of topics, including health, politics, and economics, and represents, to the best of our knowledge, the first large-scale multilingual resource for verdict generation in European languages annotated by domain experts.

We evaluated `Llama-3.1-8B-Instruct` on EuroVerdict under multiple experimental configurations, varying prompt language (English versus language-specific), training setup (zero-shot, Chain-of-Thought, and fine-tuning), and article language. Regarding the latter, we compared models conditioned on fact-checking articles in their original language against those conditioned on automatically translated versions, with a twofold motivation: to assess the model's ability to generalise across languages and evaluate its performance on non-English content, and to address the realistic scenario in which the only available evidence for a given claim is expressed in a different language. Our results demonstrate that fine-tuning consistently yields the best performance across all settings and languages, while Chain-of-Thought prompting underperforms relative to simpler baselines. Crucially, the performance gap between models conditioned on original-language articles and those conditioned on translated articles remains modest, confirming that multilingual LLMs are sufficiently robust to transfer knowledge across languages for verdict generation. Human evaluation by professional fact-checkers corroborated these findings. Taken together, these results provide a positive answer to both RQ3.1 and RQ3.2: knowledge-driven approaches can generate reliable verdicts in multilingual scenarios, even under cross-lingual knowledge conditions.

Across the chapters presented so far, a common assumption has been that reliable knowledge in the form of fact-checking articles is always available and directly paired with the claim under scrutiny. While this assumption enabled systematic benchmarking, it does not reflect the conditions under which real-world fact-checking systems must operate. In practice, a relevant fact-checking article may exist but not yet be associated with the claim at hand, or no suitable knowledge source may exist at all. Addressing these more realistic and challenging scenarios, in which knowledge retrieval and knowledge absence become central concerns, is the focus of the following chapter.

6

RAG-Based Verdict Generation: Retrieval Strategies under Claim Style Variation

6.1 Introduction

Across the preceding three chapters, knowledge-driven generation approaches were evaluated under a progressively relaxed set of assumptions. Chapter 3 established that summarisation-based pipelines can produce high-quality verdicts, demonstrating that structured knowledge extraction from fact-checking articles is a valuable step prior to generation. Chapter 4 extended this work by showing that such pipelines can further adapt their output style to social media registers, provided that a fact-checking article is already paired with the input claim. Chapter 5 relaxed the monolingual constraint, demonstrating that instruction-based LLMs can generate reliable verdicts across eight European languages, including cross-lingual configurations where the supporting knowledge is not available in the language of the claim. Yet in all three settings, one fundamental assumption remained intact: that relevant, high-quality knowledge exists and is already associated with the claim under examination.

This assumption rarely holds in practice. In real-world fact-checking workflows, claims are seldom pre-paired with a fact-checking article; more commonly, a relevant article must be retrieved from a broader knowledge base, or may not yet exist at all. This chapter directly address this problem by investigating **RQ4** through a systematic evaluation of Retrieval-Augmented Generation (RAG) pipelines for verdict generation. Specifically, we address two progressively more realistic sub-questions: how to generate high-quality verdicts when a fact-checking article exists but is not directly paired with the claim (**RQ4.1**), and how to do so when no claim-specific fact-checking article exists and the system must rely exclusively on general

reliable evidence retrieved from broader sources (**RQ4.2**).

A further dimension of realism concerns the communicative style of the input claim. While previous chapters treated claim style as a variable governing output adaptation, its impact on retrieval has not been examined. Social media claims, with their informal register, emotional content, and extra noisy information around the fact, may pose distinct challenges for retrieval systems compared to neutral, journalistic claims. This chapter, therefore, also investigates whether claim style affects RAG pipeline performance (**RQ4.3**), comparing neutral journalistic claims against social media-style claims drawn from the VerMouth dataset (see Chapter 4), which may further comprise an emotional component.

To address these questions, we present a comprehensive evaluation of RAG-based verdict generation pipelines across multiple configurations, varying retrieval strategy (sparse, dense, hybrid, and LLM-based), knowledge base composition (gold and silver), claim pre-processing (with and without fact extraction), document storage granularity (full articles versus chunks), and generation setup (zero-shot, one-shot, and fine-tuned LLMs of varying sizes). This design allows us to disentangle the contributions of individual pipeline components while assessing their interactions in increasingly realistic scenarios.

Beyond the core RAG pipelines' evaluation, this chapter also addresses the practical challenge of constructing the training data that such pipelines require. Specifically, in the last part of the Chapter, we present **First-AID**, a human-in-the-loop framework for the collection of synthetic, knowledge-grounded dialogues, and evaluated three data collection strategies that differ in the degree of human control and LLM involvement. This component of the chapter responds directly to the broader thesis objective of developing knowledge-driven generative systems that operate under realistic conditions, extending the scope of **RQ4** to include the upstream question of *how* high-quality grounded training data can be efficiently produced.

6.2 Related Work

While summarisation and instruction-based approaches, discussed in the previous chapters, can provide readable verdicts, they assume that trustworthy evidence is always available for the claim under inspection. Therefore, RAG-based approaches have been employed to guide the generation of reliable verdicts grounded in trustworthy evidence retrieved from a knowledge base (KB).

Zeng and Gao (2024) proposed JustiLM, a few-shot RAG-based approach for justification generation over real-world claims. To this end, they introduced ExClaim, a dedicated benchmark dataset, and designed a pipeline in which fact-checking articles serve as an auxiliary resource exclusively during training: at inference time, the model retrieves and conditions on general evidence rather than gold fact-checks, moving a step closer to realistic deployment conditions. Yao et al. (2023) extended RAG-based fact-checking to the multimodal setting, introducing MOCHEG, a large-scale dataset of over 15,000 claims paired with heterogeneous web evidence comprising textual paragraphs, images, videos, and tweets. Their end-to-end system jointly addresses evidence retrieval, veracity prediction, and explanation generation, providing one of the first comprehensive benchmarks for multimodal verdict production. Nevertheless, both studies concentrate on journalistic data and writing styles, with-

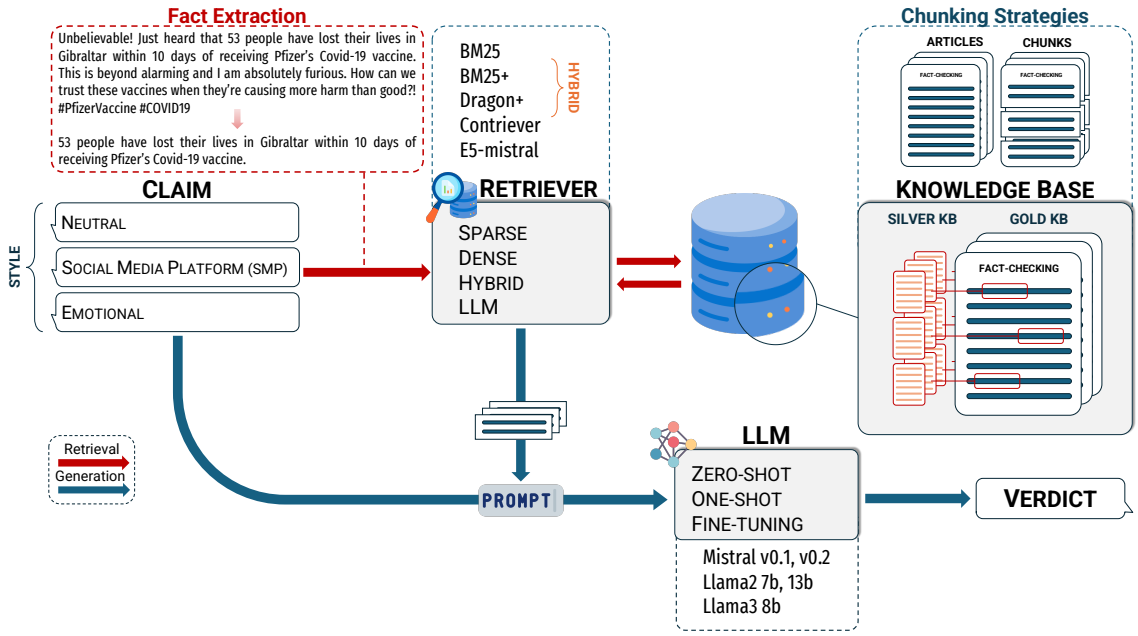


Figure 6.1: Overview of the RAG pipeline experimental design. Claims from the FullFact and VerMouth datasets are optionally preprocessed via a fact extraction module before being passed to the retrieval phase, where sparse, dense, hybrid, and LLM-based retrievers query either a Gold or Silver knowledge base. Retrieved evidence is then combined with the claim to prompt the generation module, tested in zero-shot, one-shot, and fine-tuned configurations.

out considering the communicative style employed on social media platforms, and neither examines the extent to which claim style affects retrieval performance.

Building on the work of Zeng and Gao (2024), we present an extensive evaluation of RAG pipelines for verdict generation, testing various combinations of retrieval and generation strategies. Progressing toward increasingly realistic scenarios, with respect to the workflow adopted by professional fact-checkers, we address the challenge of handling claims and sources that vary in style and complexity, aiming to closely mimic the kinds of claims that everyday users might encounter on social media platforms. Specifically, we assess the impact of claim writing style on retriever performance, highlighting where differences arise as the style shifts toward that used by SMP users. Finally, we explore the extreme scenario where no fact-checking article exists, relying solely on reliable supporting evidence.

6.3 Experimental Design

In this section, we provide details on the experimental design, as illustrated in Figure 6.1: from the datasets used, through the retrieval methods adopted, to the configurations of the LLMs employed for verdict generation.[1]

[1]Code and data are publicly available on GitHub: https://github.com/drusso98/face-the-facts

6.3.1 Dataset

To study the impact of different styles on a RAG-based verdict production task, we used the FullFact and VerMouth datasets (for details, refer to Chapters 3 and 4). The two datasets comprise eight different versions of the same claims and verdicts. FullFact provides data written in a journalistic style scraped from fullfact.org while VerMouth proposes the same data rewritten in an SMP style, and also enriched with the six emotional components defined by Ekman and Friesen (1971).

In both datasets, each claim-verdict pair is linked to a human-written fact-checking article, thus compounding to 8 different versions of the same claim: journalistic style (*neutral* hereafter), *SMP* style, anger, surprise, disgust, joy, fear, and sadness. In VerMouth, verdicts were also rewritten to reflect the various styles and emotions present in the claims. Throughout the chapter, we will refer to emotion-styled subsets as *emotional* data.[2]

6.3.2 Retrieval Module

This comprises three elements: a *query* (a claim in our case), a *knowledge base* (KB), and a *retriever*.

Claim

We used claims from FullFact and VerMouth datasets as queries. As previously described, the two datasets offer three aligned variations of a claim: neutral, SMP, and emotional. Due to *noise* in SMP and emotional data, i.e., irrelevant information surrounding the main facts, directly using claims as queries can negatively impact the retrievers' performance. *Query rewriting*, which transforms context-dependent user queries into self-contained ones, has proven to be an effective approach for enhancing retriever performance (Elgohary et al., 2019; Ye et al., 2023). For this reason, we implemented a **fact extraction module** to simplify claims and remove noise around the main fact we need to retrieve evidence for. In particular, we employed `Llama-2-13b-chat-hf` in a one-shot learning setup to extract the main facts from all SMP and emotional claims. A manual evaluation of the model's output confirmed the effectiveness of the methodology.[3] An example of an emotional claim and its related fact is provided below.

`claim`

Unbelievable! Just heard that 53 people have lost their lives in Gibraltar within 10 days of receiving Pfizer's Covid-19 vaccine. This is beyond alarming and I am absolutely furious. How can we trust these vaccines when they're causing more harm than good?! #PfizerVaccine #COVID19

53 people have lost their lives in Gibraltar within 10 days of receiving Pfizer's Covid-19 vaccine.

[2]More details on the datasets are provided in Appendix C.1.

[3]The full instruction prompt and the manual evaluation of the fact extraction module are reported in Appendix C.2.

Retriever

We evaluated several retrieval strategies, with varying computational demands to accommodate the potential computational constraints of the target users: *(i)* sparse: `BM25` and `BM25+` (Robertson et al., 1995), a popular and effective extension of tf-idf; *(ii)* dense: `Dragon+` (Lin et al., 2023) and `Contriever` (Izacard et al., 2021); *(iii)* hybrid, combining `BM25+` and `Dragon+` retrievers, using `BAAI/bge-reranker-large` as a reranker (Xiao et al., 2023); and, *(iv)* an instruction-tuned LLM for text embedding, `e5-mistral-7b-instruct` (Wang et al., 2024)[4], LLM-Retriever hereafter.

Knowledge Base

To build the KB, we employed articles from the FullFact dataset, aligned with VerMouth data. We named this KB as **Gold KB**. We experimented with two approaches: *(i)* indexing entire articles (Gold_KB$_{art}$); *(ii)* indexing small portions of each article as separate documents, i.e. *chunks*[5] (Gold_KB$_{chunks}$).

In a realistic scenario, an up-to-date KB of fact-checking articles may not be available, or a fact-checking article might not exist (yet) for a given claim. To approximate this scenario, we leveraged knowledge from reliable sources to build a **Silver KB**. Specifically, we extracted from gold fact-checking articles the evidence used to write and fact-check claims from FullFact's articles. This design choice is grounded in direct collaboration with approximately 20 professional fact-checkers. They emphasized that they do not rely on open web search, but instead consult curated and trustworthy sources retrieved via tools such as *Google Fact Check Tools* or drawn from predefined lists of reputable websites. The Silver KB thus serves as a faithful proxy for this professional workflow, making our experimental setup more realistic and practically grounded. The Silver KB was collected by following the URLs present in the articles and getting their textual content. FullFact articles also typically link to the sources of the claims. However, the reliability of these sources is questionable; thus, we filtered them out. Also, we ignored all links to social networks (Twitter, Facebook, Instagram, TikTok, and Reddit). Finally, from the remaining URLs, we extracted the text using the `Newspaper3k`[6] Python library. Statistics related to the extra evidence collection are presented in Appendix C.3.

6.3.3 Verdict Generation

For the generation of the verdicts, we tested five LLMs, selected based on differences in sizes or the presence of guardrails: `Mistral`, in its v1.0 and v2.0 versions (Jiang et al., 2023); `Llama-2` (Touvron et al., 2023), in its 7B and 13B chat versions;[7] and `Llama-3-8B-Instruct`.[8] We combined the claim and the retrieved evidence to prompt the LLM (see Appendix C.5.1), and tested generation under different setups, namely zero-shot, one-shot, and fine-tuning. For fine-tuning, we employed `Llama-2-13b`, the best-performing model in zero-shot and one-shot settings.

[4] https://hf.co//intfloat/e5-mistral-7b-instruct

[5] We used the `LlamaIndex` (Liu, 2022) sentence splitter, which minimises text fragmentation by keeping sentence integrity, with a maximum chunk token size of 100.

[6] https://github.com/codelucas/newspaper/

[7] https://hf.co/meta-llama/Llama-2-7b-chat-hf; https://hf.co/meta-llama/Llama-2-13b-chat-hf

[8] https://hf.co/meta-llama/Meta-Llama-3-8B-Instruct

		sparse	dense	hybrid	LLM-Retriever	sparse	dense	hybrid	LLM-Retriever
		hit_rate@1				**mrr@10**			
Articles	Neutral	0.903	0.905	**0.966**	0.960	0.931	0.938	0.972	**0.978**
	SMP	0.770	0.799	**0.937**	**0.937**	0.817	0.866	**0.963**	0.962
	Emotional	0.778	0.839	0.905	**0.938**	0.838	0.866	0.933	**0.964**
	SMP_{facts}	0.778	0.801	**0.937**	0.914	0.837	0.891	**0.963**	0.947
	Emotional_{facts}	0.835	0.846	0.905	**0.932**	0.883	0.897	0.933	**0.958**
		hit_rate@10				**map@10**			
Chunks	Neutral	0.963	0.992	**1.000**	**1.000**	0.392	0.552	0.573	**0.655**
	SMP	0.856	0.974	0.994	**1.000**	0.275	0.484	0.536	**0.619**
	Emotional	0.904	0.977	0.994	**1.000**	0.273	0.482	0.545	**0.599**
	SMP_{facts}	0.905	0.972	0.994	**1.000**	0.304	0.505	0.526	**0.601**
	Emotional_{facts}	0.939	0.978	0.994	**0.999**	0.345	0.518	0.552	**0.615**

Table 6.1: Results for the retrieval experiments. We report hit_rate, mrr, and map for retrieval over the Gold_KB_{art} (Articles) and the Gold_KB_{chunks} (Chunks) KBs. SMP_{facts} and Emotional_{facts} indicate input preprocessing with the fact extraction module. The first and second best results for claim style are in bold and underlined, respectively.

6.4 Retrieval Experiments

Retrieval from Gold KB

We tested the retrievers on FullFact and VerMouth test sets with an increasing number of retrieved documents ($k = 1, \ldots, 10$). For each claim used as a query, we considered as relevant documents the fact-checking article, or its chunks, linked to the claim. For space reasons, results on the emotional datasets will be presented in aggregated form, referred to as the *'emotional'* set. Experiments were carried out integrating into the LlamaIndex (Liu, 2022) framework either `Rank-BM25` (Brown, 2020) or HuggingFace's models (Wolf et al., 2020); retrieval performance was assessed with `ranx` (Bassani, 2022) using Hit Rate, Mean Reciprocal Rank (mrr), and Mean Average Precision (map) as evaluation metrics; detailed descriptions of these metrics are provided in Section 2.4.5 of Chapter 2.

Table 6.1 presents the results for each retrieval approach (sparse, dense, hybrid, LLM-Retriever) across all claim's styles (neutral, SMP, emotional) and fact-extraction pre-processings (SMP_{facts}, emotional_{facts}) using both KB configurations ($Gold\_KB_{art}$ and $Gold\_KB_{chunks}$). For $Gold\_KB_{art}$, we report hit_rate@1 and mrr@1 (Mean Reciprocal Rank), as each claim had only one gold related article. For $Gold\_KB_{chunks}$, we report hit_rate@10 and map@10 (Mean Average Precision) to assess whether the retrievers could consistently include at least one gold chunk among the top 10 and the precision of the retrieval system across different recall levels.[9]

For article retrieval (upper part of Table 6.1), all four retrievers achieved high accuracy on neutral claims, with a hit_rate@1 above 90%. When the correct article was not immediately retrieved, they still ranked it highly, as shown by strong MRR@10 scores. Performance declined with noisier claims, especially for sparse

[9]More details in Appendix C.4.2.

retrievers, while dense models and the LLM-Retriever showed greater robustness. Hybrid retrieval (combining sparse and dense) performed comparably to the LLM-Retriever.

In chunk retrieval (lower part of Table 6.1), the LLM-Reranker excelled, achieving a hit_rate@10 that always included a gold chunk and reaching an average MAP@10 of 70%. Sparse retrievers showed low map scores (∼40% for neutral claims), particularly for SMP and emotional claims (<30%), while dense and hybrid approaches followed trends similar to article retrieval. Thus, the low MAP scores indicate that sparse retrievers must retrieve more chunks from the knowledge base to select the relevant content. However, this comes at the cost of also retrieving more non-relevant content, which could potentially compromise the subsequent generation phase.

Overall, the LLM-Retriever consistently outperforms other approaches. Notably, it remains stable even when exposed to different input claim styles, with minimal degradation when noise is introduced. A paired t-test confirms that the performance gains of LLM-Retrieval are statistically significant compared to other methods, with the exception of the hybrid retriever over hit_rate@1. Still, it maintains a slightly higher mean score (0.934 vs. 0.925) and lower variance (0.061 vs. 0.069). Dense retrievers perform worse but show robustness to stylistic variations. In contrast, sparse retrievers are significantly affected by data noise, resulting in performance drops across all three datasets: we find that the fact extraction module we included yields consistent performance improvements, and particularly helps when using sparse and dense retrievers.

Retrieval from Silver KB

We tested the optimal retriever methodologies from the previous experiments, specifically the LLM-Retriever and the hybrid retriever. As outlined in Section 6.3.2, the Silver KB consists of reliable sources extracted from the initial fact-checking articles, amounting to 9983 chunks in total. Since no gold fact-checking article is directly available in this setting, the corresponding FullFact article was retained as the relevance ground truth for evaluation purposes only.

The results (Table 6.2) show that modifying the knowledge base strongly impacted retrieval performance both across the three datasets and the two retrieval strategies. In particular, the two retrievers exhibited comparable performance, mirroring the behaviour observed in earlier experiments with the Gold KB. Unlike the previous setting with the Gold KB, the LLM-Retrieval's performance in this context is markedly influenced by the stylistic nature of the claims: neutral formulations consistently yielded higher results, and performance was impacted by the claims' complexity.

Prepending a fact extraction module generally improves retrieval results: even robust retrievers can benefit from preprocessing when dealing with heterogeneous KB (that do not contain gold fact-checking articles) and complex (e.g., emotional) claims.

	Neutral	SMP	Emotional	SMP_{facts}	Emotional_{facts}
LLM-Retriever	0.683	0.652	0.637	0.689	0.671
Hybrid	0.683	0.652	0.631	0.602	0.652

Table 6.2: Hit_rate@10 scores for retrieval with LLM-Retrieval and hybrid retriever over the Silver KB.

6.5 Generation Experiments

For verdict generation, the claim and its corresponding retrieved evidence were combined into a prompt fed to the five different LLMs (Section 6.3.3). For evidence retrieval, we employed the best-performing retriever (i.e., LLM-Retriever). We tested the five LLMs under three setups: zero-shot, one-shot, and fine-tuned.[10] For fine-tuning, we employed the best-performing model in both the zero-shot and one-shot configurations, `Llama-2-13b`.

When using the Gold_KB_{art}, we included the top-1 article (i.e., the most relevant) in the prompt, a choice justified by the LLM retriever's remarkable hit_rate@1 results. Conversely, with retrieval from Gold_KB_{chunks}, we fed the model with 10 retrieved chunks, based on its map@10 performance (see Table 6.1 and Figure C.4.2).[11]

Automatic Metrics

Inspired by previous works (e.g., Zeng and Gao, 2024), we use ROUGE-LSum (Lin, 2004) (*RL-Sum*), BARTScore (Yuan et al., 2021), and SummaC (Laban et al., 2022) to evaluate lexical adherence and faithfulness of the generated text to the context provided to the LLM. Further, we used BARTScore, which is unaffected by differences in text length, to compute the semantic similarity (*GoldSim*) between the generated and gold verdicts from FullFact and VerMouth. A detailed description of the metrics employed is provided in Section 2.4.5 of Chapter 2. For average performances per dataset and per model, see Tables 6.3 and 6.4, respectively.[12]

6.5.1 Zero-shot and One-shot

Generations with neutral claims yielded better results compared to the SMP and emotional data in both zero-shot and one-shot experiments, indicating that the complexity of claims affects not only the retrieval phase but also the generation phase. Interestingly, when comparing the similarity between the generated and the gold verdicts, the generations with emotional data produced the best results (Table 6.3). Upon manual inspection, we found recurrent patterns in the SMP and emotional data, such as expressions of empathy and politeness typical of ChatGPT that was used to generate the manually curated verdicts of VerMouth ("I understand your frustration", "It is important to note that"). These patterns were also replicated by the models used in this study. Notably, zero-shot experiments generally outperformed one-shot experiments when results were averaged across all the data (Table 6.3) and LLMs (Table 6.4). Turning to individual model performances

[10]The instructions used are provided in Appendix C.5.1.
[11]Examples of generations are provided in Appendix C.5.5.
[12]Complete results can be found in Appendix C.5.4, Table C.4.

(Table 6.4), larger models (`Llama-2-13b`) demonstrated higher faithfulness to the context and similarity to gold verdicts, whereas smaller ones (`mistral-7b-v0.1`, `llama3-8b`) showed better contextual adherence in terms of overlaps (R-LSum) and consistency (SummaC). Misalignments between SummaC and GoldSim were observed, often stemming from the fact that fact-checking articles might contain

Original

`claim`

Face masks can cause low oxygen levels in the body.

`verdict`

No, the evidence suggests that masks do not cause low oxygen levels.

SMP-style

`claim`

#StaySafe alert: Face coverings can lead to decreased O2 in the body, so watch for any signs of trouble breathing. Wear masks only if necessary & keep an eye out for signs of low oxygen. #Masks #OxygenLevels #Health

`verdict`

Thank you for sharing this information. It's important to note that the World Health Organization (WHO) and other medical professionals have consistently stated that face masks are a safe and effective way to reduce the transmission of COVID-19. While it's possible that masks may cause some discomfort or difficulty breathing, there is no evidence to suggest that they can cause a significant decrease in oxygen levels. If you have any concerns about your health or the health of others, it's always best to consult with a medical professional. #StaySafe #FaceMasks #COVID19

Emotional style

`claim`

I'm really scared about the thought of wearing a face mask causing low oxygen levels in my body. I know it's important to protect others, but I don't want to harm myself in the process. #COVID19 #facemasks #safetyconcerns

`verdict`

I understand your concern about wearing face masks and the potential for low oxygen levels. However, there is no evidence to suggest that face masks cause low oxygen levels. In fact, many fact-checkers have debunked this claim. It's important to prioritize your safety while protecting others, and face masks are a crucial tool in doing so. #COVID19 #facemasks #safety

Example 6.1: Example of generations using claims form FullFact and VerMourt (first line), e5-mistral as a retriever, and `LLaMA-2-13b` fine-tuned for the generation of the verdict (second line).

multiple supporting arguments. When the retriever or LLM selects only a subset, the generated verdict may diverge from the gold verdict in argumentation while remaining contextually accurate.

To sum up, generations with neutral claims outperformed those with emotional data in both zero-shot and one-shot experiments but produced more accurate results when paired with emotional data. In terms of faithfulness to context and similarity to the gold, larger models generally performed better. However, smaller models exhibited superior contextual adherence.

6.5.2 LLM Fine-Tuning

We fine-tuned `Llama-2-13b`, the best-performing model in the previous in-context learning experiments, disjointly on the three claim styles. To this end, the model was fed with claims and gold verdicts from the FullFact and VerMouth datasets, complemented by positive and negative contextual information.[13] We randomly sampled 200 entries from both neutral and SMP training datasets; similarly, to obtain a comparably sized dataset for emotional data, we randomly sampled 35 examples for each of the 6 emotional dimensions available.[14] Examples of generation are reported in Table 6.1. Fine-tuning results (Tables 6.3 and 6.4) show that `LLaMA-2-13b` models improve in faithfulness and similarity w.r.t. gold verdicts across all datasets and for both full articles and text chunks. However, this comes at the cost of lower ROUGE-L scores: without fine-tuning, the models tend to extract and replicate not only the necessary information but also the exact wording from the context. Thus, fine-tuned models abstract better from the context, as expected, and also show better performance in selecting relevant and reliable information. Further, after fine-tuning, the emotional models achieved higher similarity scores to the original claims. Akin to the zero and one-shot configurations, inspection of the generated verdicts reveals that these models produce empathetic expressions similar to those found in VerMouth.

		Articles				Chunks			
		R-LSum	**BARTScore**	**SummaC**	**GoldSim**	**R-LSum**	**BARTScore**	**SummaC**	**GoldSim**
zero shot	Neutral	**0.16**	**-2.13**	**0.35**	-3.03	**0.25**	**-2.61**	**0.25**	-3.06
	SMP	0.15	-2.31	0.32	-3.01	0.24	-2.69	0.24	-3.03
	emotional	0.14	-2.48	0.31	**-2.94**	0.22	-2.79	0.23	**-2.98**
one shot	Neutral	**0.16**	**-2.13**	**0.33**	-3.03	**0.26**	**-2.53**	0.24	-2.99
	SMP	0.15	-2.42	0.32	-3.01	0.24	-2.73	**0.23**	-2.95
	emotional	0.14	-2.60	0.32	**-2.95**	0.22	-2.85	**0.23**	**-2.91**
fine tuning	Neutral	**0.10**	**-1.45**	**0.53**	-2.71	0.08	**-1.42**	**0.52**	-2.75
	SMP	**0.10**	-2.30	0.33	-2.58	**0.10**	-2.45	0.32	**-2.63**
	emotional	**0.10**	-2.43	0.31	**-2.48**	**0.10**	-2.76	0.31	-2.68

Table 6.3: Generation results per dataset, *averaged across the LLMs*. Retrieved articles or chunks were employed in the generation. The best results for each generation configuration are in bold.

[13] Details are provided in Appendix C.5.2.

[14] Fine-tuning details are reported in Appendix C.5.3.

		Articles				Chunks			
		R-LSum	**BARTScore**	**SummaC**	**GoldSim**	**R-LSum**	**BARTScore**	**SummaC**	**GoldSim**
zero-shot	mistral-v0.1	**0.19**	-2.16	**0.35**	-3.02	**0.19**	-2.20	**0.35**	-3.05
	mistral-v0.2	0.13	-2.55	0.32	-3.12	0.15	-2.50	0.33	-3.11
	llama3-8b	0.05	-3.21	0.30	-3.54	0.04	-3.37	0.32	-3.65
	llama2-7b	0.16	-2.12	0.32	-2.84	**0.19**	-2.06	0.33	-2.87
	llama2-13b	0.17	**-1.89**	0.34	**-2.73**	**0.19**	**-1.89**	**0.35**	**-2.77**
one-shot	mistral-v0.1	**0.16**	-2.38	**0.33**	-3.02	0.17	-2.47	0.32	-3.07
	mistral-v0.2	0.12	-2.61	0.31	-3.16	0.13	-2.63	0.36	-3.16
	llama3-8b	0.13	**-2.25**	**0.33**	-2.98	0.14	-2.26	**0.33**	-2.97
	llama2-7b	**0.16**	-2.35	0.32	-2.93	**0.19**	-2.34	0.31	-2.89
	llama2-13b	**0.16**	-2.31	0.32	**-2.90**	0.17	**-2.22**	0.31	**-2.79**
ft	llama2-13b	0.10	-2.08	0.39	-2.59	0.10	-2.21	0.38	-2.69

Table 6.4: Generation results per LLM, *averaged across the three datasets* (neutral, SMP, emotional). Retrieved articles or chunks were employed in the generation. The best results for each generation configuration are in bold.

6.5.3 Generation with Silver KB

Finally, we tackled the scenario where the useful information is spread across several documents: this lifts the constraint of existing datasets wherein a claim is paired to a single article. We used all documents from the Silver KB (Section 6.3.2) and the LLM-Retriever/`Llama-2-13b` setup (best-performing in the above). We focused solely on the chunk-based configuration as the information required to build a verdict is *(i)* often distributed across multiple extra documents, and *(ii)* it is more likely to be located in specific sections of these extra evidence articles. Results (Table 6.5) show that, except for R-LSum, `Llama-2-13b`'s performance is consistently slightly worse when compared to generation using the Gold KB.[15] Still, a qualitative analysis showed that using the Silver KB resulted in verdicts that, in most cases, were consistent with the claim, faithful to the context, and informative.

The lower results can be explained by the substantial difference between the Gold and the Silver KBs: a gold fact-checking article refers to a single claim and contains all the information needed to generate the verdict. Therefore, when using the Gold KB, out of a total of ten chunks, a robust retriever – such as LLM-Retriever – can identify a larger number of informative chunks. Conversely, in the case of the Silver KB, for each claim, on average, there are four related articles (see Appendix C.3, Table C.2) that are most likely to provide partial information about the verdict. Therefore, realistic retrieval scenarios for professional fact-checkers involve large, informationally sparse, and repetitive document collections, meaning 10 retrieved chunks may lack sufficient information for a good verdict.

6.5.4 Human Evaluation

Automatic metrics for NLG evaluation are known to correlate only partially with human judgments. Several works showed how optimizing for such metrics (e.g., ROUGE) is largely suboptimal (Paulus et al., 2018; Scialom et al., 2019), and they suffer from weak interpretability and failure to capture nuances (Sai et al., 2022). Therefore, we also provide a comprehensive human evaluation of the generated ver-

[15] Compare with Tables 6.4, 6.3, and Appendix C.5.4 - Table C.4.

		R-LSum	BARTScore	SummaC	GoldSim
zero-shot	Neutral	0.23	-2.17	0.28	-3.04
	SMP	0.21	-2.30	0.28	-2.84
	Emotional	0.19	-2.50	0.26	-2.73
	Mean	0.21	-2.32	0.28	-2.87
one-shot	Neutral	0.24	-2.39	0.24	-3.02
	SMP	0.21	-2.57	0.25	-2.90
	Emotional	0.18	-2.68	0.26	-2.76
	Mean	0.21	-2.55	0.25	-2.89
fine tuned	Neutral	0.12	-2.18	0.39	-3.00
	SMP	0.13	-2.64	0.26	-2.59
	Emotional	0.13	-2.97	0.25	-2.70
	Mean	0.13	-2.60	0.30	-2.77

Table 6.5: Generation results with Silver KB.

dicts.

We enlisted three expert evaluators,[16] and we focused on the data generated using Llama2-13b and the LLM-Retrieval over the Gold KB, as this configuration yielded the best results in both the retrieval and generation phases (see Sections 6.4, 6.5.3). The evaluators were provided with pairs of verdicts (either gold or generated using zero-shot, one-shot, or fine-tuned models) and their corresponding claims. They were instructed to assess the best verdict based on three aspects: effectiveness, informativeness, and emotional/empathetic coverage. We sampled 72 claims equally distributed among the test sets, and provided the evaluators with six combinations of verdict pairs, compounding to 432 samples to be evaluated on the three evaluation dimensions; thus, the reported human evaluation is based on a total of 1296 data points. In Figure 6.2 (left), we report how many times, in percentage, the human annotators preferred each of the four verdict generation setups (gold, zero-shot, one-shot, fine-tuning) across the three claim styles (neutral, SMP, emotional). The interannotator agreement (Cohen's κ) was 0.7, indicating good agreement.

Results show that zero-shot and one-shot approaches were largely preferred in terms of informativeness. Gold and fine-tuned configurations were considered the best in terms of emotional matching between the claim and the related verdicts, which can be explained by the fact that fine-tuned models learned to mimic the emotional style of the gold training data. This is supported by the effectiveness evaluation in Figure 6.2 (right), where the delta between the preferences for zero/one-shot configurations and gold/fine-tuned ones is higher for neutral claims, and it decreases as the emotional component increases. Nevertheless, on average, zero and one-shot configurations remain the most preferred. This result could stem from FullFact's verdicts' intended use alongside articles, not as standalone social corrections, and

[16] A senior researcher and two MSc graduates; all volunteer evaluators are proficient in English, experts in NLP, and knowledgeable about social media platforms' communication styles and dynamics, particularly in the context of misinformation.

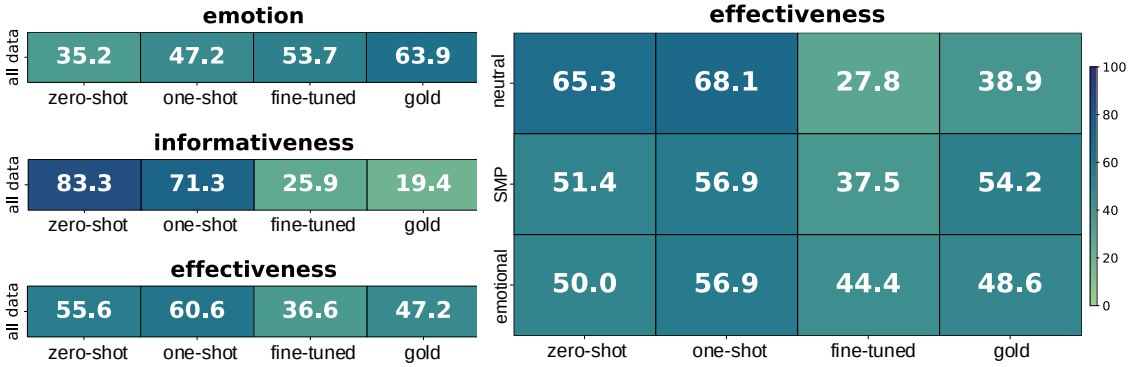


Figure 6.2: Human evaluation results: Percentages of preference for the four generation setups across the datasets.

consequently aren't completely self-contained in grounding, with the full context available on the article page. [17] Finally, a fine-grained analysis of annotator agreement across the three aspects revealed a high κ score for informativeness (0.8), and moderately strong scores for the more subjective dimensions of emotional coverage and effectiveness (both at 0.6).

6.6 Improving Data Collection for RAG-based Dialogue Generation

The experimental results presented in this chapter establish that fine-tuning on task-specific, knowledge-grounded data constitutes an effective strategy for improving the faithfulness and stylistic appropriateness of generated verdicts in RAG-based pipelines. This finding, however, raises a fundamental methodological question that extends beyond the scope of verdict generation: how can sufficiently large and high-quality corpora of knowledge-grounded dialogues be constructed to support the fine-tuning of generative models in domains where expert annotation is scarce and costly? This question is particularly pressing in the context of knowledge-driven counter-dialogue generation, where factual faithfulness to a reference document must be ensured at every turn of the conversation, a constraint that neither fully automated LLM-based generation nor purely manual annotation can reliably satisfy at scale.

To address this challenge, we developed **First-AID** (First Annotation Interface for grounded Dialogues), a human-in-the-loop (HITL) data collection framework for the knowledge-driven generation of synthetic dialogues grounded in textual documents. The framework targets the creation of RAG-structured dialogue data in which each conversational turn is explicitly linked to one or more passages from a reference document, enabling the resulting corpora to be used not only for training dialogue models but also for training and evaluating the retrieval components of RAG-based pipelines. The framework was evaluated on data collection for misinformation and hate speech countering, two domains that are central to this thesis and that exemplify the broader challenges of knowledge-driven counter-narrative generation.

[17] Further details are provided in Appendix C.6.

Data Collection Strategies

A central contribution of First-AID is its support for three distinct data collection strategies, each representing a different position on the automation-control spectrum (see Figure 6.3). Rather than committing to a single workflow, the framework is designed to allow practitioners to select the strategy most appropriate to their specific constraints:

1. **Manual**: the annotator writes the dialogue from scratch based on one or more reference documents, manually linking each turn to the relevant textual spans. This strategy maximises annotator control and minimises LLM-induced biases, at the cost of higher annotation effort.
2. **Pre-compiled**: an LLM automatically generates a complete dialogue grounded in the provided documents, with turns paired to source passages. The annotator subsequently reviews and post-edits both the dialogue content and the grounding links. This strategy prioritises throughput, but may result in reduced turn diversity.
3. **Interactive**: the annotator is assisted turn-by-turn by an LLM, which proposes multiple candidate continuations at each dialogue step. The annotator selects and optionally revises one of the generated options, with each choice conditioning the generation of subsequent turns. This strategy is designed to balance annotator agency with annotation efficiency.

This modular architecture reflects a broader design principle: that the appropriate degree of human intervention in knowledge-driven data collection is not fixed, but depends on the relative weight assigned to data quality, annotator expertise, and resource constraints in a given scenario. The platform further supports multiple annotation layers, enabling expert curators to post-edit dialogues produced by less experienced annotators, as well as iterative model improvement cycles in which post-edited outputs are incorporated into subsequent training iterations.

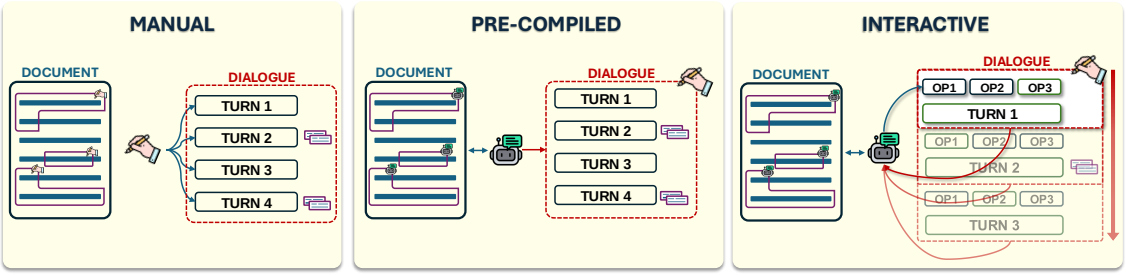


Figure 6.3: Graphical representation of the three data collection strategies supported by First-AID platform.

The Annotation Interface

The annotation interface operationalises the RAG-oriented design of the framework by structuring the annotation task around explicit document grounding. The interface is organized into three panels: a document viewer on the left, where annotators can highlight and assign text spans to individual dialogue turns; a central dialogue

editor, where turns can be written or revised; and a right panel listing the grounding spans linked to each turn (See Image 6.4). This layout enforces grounding at the turn level rather than at the dialogue level, a property that is essential for training retrieval components and rerankers in RAG-based systems. The platform connects to a customizable API, allowing practitioners to configure the LLMs and retrievers employed during generation, as well as the communicative roles and styles of the dialogue participants. The software is implemented using VueJS (frontend) and Python/SQLite (backend) and is publicly released as an open-source package.[18]

Evaluation

The platform was evaluated through four annotation sessions involving 50 domain experts on a data-curation task of misinformation and hate countering dialogues rooted in fact-checking articles. Experts comprised professional fact-checkers from recognized organizations and NGO members working on hate speech countering, drawn from four European countries. Each session combined a hands-on annotation phase using all three collection strategies with structured qualitative interviews and quantitative analysis of activity logs.

Quantitative results, summarized in Table 6.6, reveal a clear efficiency advantage for the *interactive* strategy, which achieves an average annotation rate of 17.68 words per minute and a mean dialogue completion time of 479 seconds, compared to 825 seconds for the pre-compiled and 1006 seconds for the manual strategy. The *pre-compiled* strategy produces the longest dialogues on average (6.25 turns, 162 words per dialogue), suggesting that unconstrained LLM generation favours elaboration over conciseness. Grounding coverage remains consistently high across all modalities, with 70–86% of turns linked to at least one source passage, confirming that the platform is effective at producing the turn-level grounding annotations required for

[18] https://github.com/LanD-FBK/first-AID

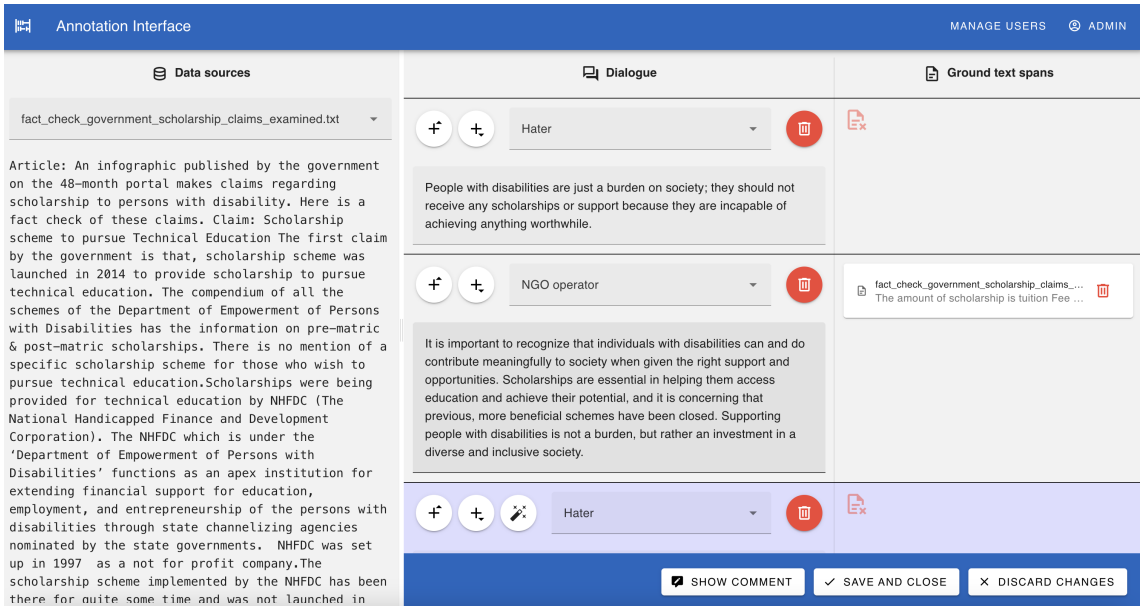


Figure 6.4: First-AID annotation screen.

	Avg. Time per Dialogue (sec)	Avg. Turns per Dialogue	Avg. Words per Dialogue	Words per minute	Turns per minute	Turns with Ground (%)	Avg. Number of Grounds per Turn
Manual	1006.06	4.74	132.79	7.92	0.28	70.90	1.56
Pre-compiled	825.37	6.25	162.28	11.80	0.45	83.15	2.18
Interactive	479.28	3.98	141.26	17.68	0.50	85.88	1.10

Table 6.6: Dialogues statistics over the data collected through the three different collection strategies.

RAG-based training.

Qualitatively, annotators evaluated the *interactive* strategy as the most effective balance between efficiency and control, characterising it as facilitating the annotation process while preserving human oversight. The *pre-compiled* strategy was valued for its speed but criticized for producing repetitive outputs that struggled with nuanced linguistic phenomena. The *manual* strategy, while the most demanding in terms of effort, was commended for producing dialogues free of LLM-induced artifacts. Taken together, these results suggest that no single strategy is universally optimal; the choice among them should be governed by an explicit analysis of the trade-offs between annotator control, output diversity, and throughput in a given data collection context.

6.7 Conclusions

This chapter addressed **RQ4** and its sub-questions by providing a comprehensive evaluation of RAG-based pipelines for verdict generation under increasingly realistic fact-checking conditions. Across the preceding chapters, knowledge-driven generation approaches had consistently operated under the assumption that relevant, high-quality evidence is directly paired with the input claim. This chapter challenged that assumption by systematically examining what happens when it no longer holds. To this end, we evaluated a wide range of RAG pipeline configurations, varying retrieval strategy (sparse, dense, hybrid, and LLM-based), knowledge base composition (gold and silver), document storage granularity (full articles versus chunks), claim preprocessing (with and without fact extraction), and generation setup (zero-shot, one-shot, and fine-tuned LLMs of varying sizes). This design allowed us to disentangle the contribution of individual pipeline components while assessing their interactions across scenarios of increasing realism, directly addressing **RQ4.1** and **RQ4.2**.

Our results show that LLM-based retrievers consistently outperform other methods in terms of adaptability, though they face challenges with heterogeneous knowledge bases. Hybrid retrieval offers a computationally efficient alternative, while dense retrievers exhibit robustness to stylistic variation. Fact extraction modules improve retrieval performance across all configurations, confirming the value of query preprocessing for handling informal or emotionally charged claims, as examined under **RQ4.3**. On the generation side, larger models produce more faithful and contextually aligned verdicts, while human evaluators consistently preferred zero-shot and one-shot outputs for their informativeness. Taken together, these results provide a positive answer to **RQ4**: RAG-based pipelines are effective for knowledge-driven verdict generation in realistic settings, even when no gold fact-checking article is directly available.

The chapter further introduced First-AID, a human-in-the-loop annotation framework for the creation of knowledge-grounded dialogue corpora. By supporting three distinct data collection strategies spanning a spectrum from fully manual to largely automated, First-AID addresses the data scarcity that constrains the fine-tuning of dialogue-oriented RAG systems. Its evaluation with domain experts confirmed the interactive strategy as the most effective trade-off between annotation efficiency and output quality, with consistently high grounding coverage across all modalities.

Throughout this and the preceding chapters, misinformation has been treated as a self-contained textual unit, here referred to as a *claim*: a discrete, well-formed statement whose veracity can be assessed, and which can be directly fed into the knowledge-driven pipelines described so far. However, misinformation does not always manifest in this canonical form. In the next chapter, we turn to a more subtle and pervasive variant: *clickbait*. Clickbait headlines exploit the strategic omission of information as a curiosity cue, drawing reader attention through what is left unsaid rather than through explicit false assertion. As such, they cannot be directly inserted into the knowledge-driven pipelines developed in this thesis without a prior preprocessing step that reinstates the omitted factual content. The following chapter investigates how knowledge-driven generation systems can be employed precisely for this preprocessing step, transforming informationally incomplete clickbait headlines into factually grounded claims suitable for downstream verification.

7

Incomplete Claims: Clickbait Detection, Spoiler Generation, and Neutralisation

7.1 Introduction

Across the preceding four chapters, knowledge-driven generative approaches were systematically evaluated under progressively relaxed assumptions. Chapter 3 and Chapter 4 demonstrated that summarisation-based pipelines can produce high-quality verdicts and adapt to both journalistic and social media communication styles, provided a fact-checking article is paired with the input claim. Chapter 5 relaxed the monolingual constraint, showing that instruction-based LLMs can generate reliable verdicts across multiple European languages, including cross-lingual settings. Chapter 6 addressed the more realistic scenario in which knowledge is not directly paired with the claim or may not yet exist, proposing RAG-based pipelines that retrieve and reason over heterogeneous evidence sources. Across all four settings, however, one fundamental assumption remained constant: misinformation was treated as a discrete, verifiable claim that can be directly submitted to a fact-checking pipeline.

This assumption does not hold universally. Not all misinformation takes the form of an explicit false statement amenable to direct verification. As noted in Chapter 6, incorporating preprocessing modules for fact extraction was shown to improve retrieval performance when dealing with complex, noisy, or stylistically challenging claims. This observation points to a broader challenge: when the input itself is not a well-formed, verifiable claim, knowledge-driven generation systems cannot be applied directly. Instead, a preprocessing step is required to transform the input into a suitable form before it can be fed into the generative pipelines developed in previous chapters.

Clickbait headlines exemplify this challenge. Unlike standard misinformation claims, clickbait headlines mislead readers through sensationalism, curiosity gaps, and deliberate vagueness rather than through verifiable falsehoods. The deception lies not in a false statement but in the mismatch between the expectations raised by the headline and the content delivered by the associated article. As a consequence, the conventional paradigm of matching a claim to a fact-checking article does not apply: before any knowledge-driven verdict generation can occur, the headline must first undergo *factuality completion*, a preprocessing step that leverages the knowledge embedded in the source article to reconstruct a full, verifiable claim from the incomplete or misleading headline.

This chapter directly addresses this challenge by investigating **RQ5**: *How can knowledge-driven approaches address complex misinformation that cannot be countered through traditional claim-based fact-checking?* Specifically, we examine how the information already present in the misinformation itself, namely the clickbait article body, can be exploited as a source of in-domain knowledge to neutralise the misleading headline. In doing so, we demonstrate how knowledge-driven generation techniques can serve not only as verdict generation systems but also as input preprocessing tools that make non-standard misinformation amenable to the pipelines developed in previous chapters.

To investigate this question, we focus on clickbait, a setting that has received limited attention in the literature. To this end, we introduce *ClickBaIT*, a novel Italian dataset of manually annotated clickbait articles, which supports three interconnected tasks: clickbait *detection*, *spoiler generation*, and a novel task we propose, *clickbait neutralisation*[1]. The latter consists of rewriting the clickbait headline into a factually accurate and informative version by leveraging the spoiler, which is itself generated from the article body. This pipeline (depicted in Figure 7.1) constitutes a concrete instantiation of knowledge-driven preprocessing for non-standard misinformation: the knowledge resides in the misinformation artefact itself, and generative models are employed to extract and restructure it into a form that can be countered or used to feed downstream verdict generation systems.

7.2 Related Work

Accuracy and truthfulness are essential characteristics of journalism. Nevertheless, in an effort to improve revenue, a large number of newspapers and magazines publish *clickbait* articles, a viral journalism strategy that seeks to attract users to click on a link to a page through tactics such as sensationalist stories and catchy headlines that act as bait. The use of these tactics harms the quality of news pieces and thus hinders the ability of citizens to obtain reliable and objective information. The literature distinguishes between two main types of clickbait. (i) *Classical clickbait* (Scott, 2021) embeds within the headlines information gaps, also known as curiosity gaps (Blom and Hansen, 2015; Loewenstein, 1994), in order to arouse curiosity in the reader that is forced to access the article's content which is ultimately disappointing. Classical clickbait usually makes use of hyperbolic language, caps lock, demonstrative pronouns and superlative to grasp the user's attention (Scott, 2021;

[1]The dataset is available in https://github.com/oaraque/ClickBaIT

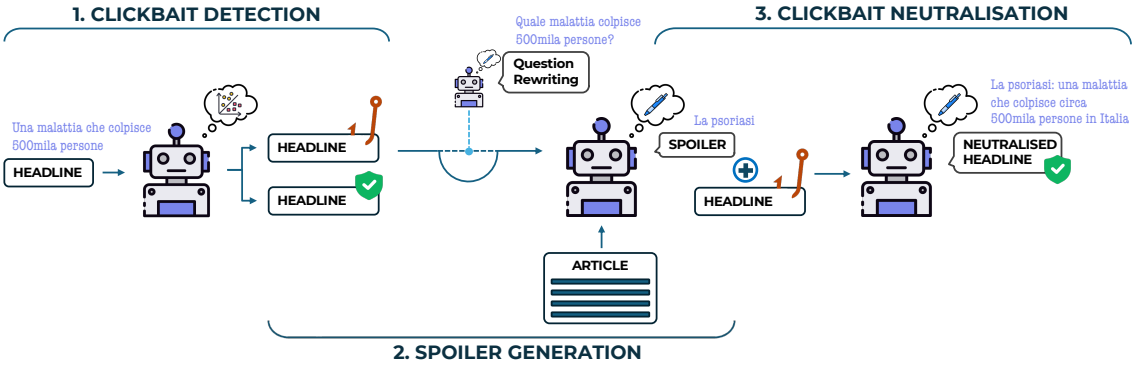


Figure 7.1: The experimental design is depicted, encompassing three tasks: *clickbait detection*, *spoiler generation*, and *clickbait neutralisation*. The robot icon represents the language model used for either classification or generation. We utilized *DistilBERT* and *Llama3-8B* for task 1, and *LLaMAntino-3-8B* for tasks 2 and 3. The models were tested for generative tasks using zero-shot, few-shot, and fine-tuning configurations, except for question rewriting, for which we employed a few-shot approach.

Scott and Jackson, 2020; Scott, 2023). (ii) *Deceptive clickbait* (Scott, 2023) refers to headlines that resemble traditional media headlines by offering a summary of the article, still leading to content that differs from the reader's expectations. These headlines promise high news value but deliver content with low news value, resulting in reader disappointment.

Although clickbait headlines are considered one of the less harmful forms of fake news, as their main goal is to increase profit by driving traffic to their website (Zannettou et al., 2019; Aïmeur et al., 2023), they can sometimes pose a danger, especially when they deal with potentially harmful topics such as health and science. To address this problem, Natural Language Processing techniques have been widely employed to detect clickbait headlines, with a particular focus on the English language (Potthast et al., 2016; Rajapaksha et al., 2021). Hagen et al. (2022) proposed the clickbait spoiling task, i.e., the generation of a short text that satisfies the curiosity induced by a clickbait post.

Several works address clickbait detection: Potthast et al. (2016) collected a corpus of clickbait articles, posted by well-known English-speaking newspapers on Twitter, and proposed a set of lexical and semantic features to be used with a Random Forest classifier. Following the general trend in Natural Language Processing (NLP) field, clickbait detection has also been explored using deep learning methods, such as convolutional (Agrawal, 2016) and recurrent (Kaur et al., 2020) neural networks, as well as more recent Transformer-based approaches (Rajapaksha et al., 2021).

Other works leveraged Natural Language Generations (NLG) strategies to create a piece of text, the *spoiler*, comprising the information needed to fulfil the curiosity gap present in clickbait headlines. This task was proposed by Fröbe et al. (2023) with the name of *spoiling generation*. The authors created the *Webis Clickbait Spoiling Corpus 2022*, and cast spoiler generation as a Question Answering task. Eventually, they open the challenge to the community through a SemEval-2023 shared task (Fröbe et al., 2023; Ojha et al., 2023). The optimal spoiler generator operates with

five independent sequence-to-sequence generative models. It selects the best spoiler through a majority vote, determined by comparing edit distances among the outputs (Kurita et al., 2023).

Regarding the languages studied, the majority of works are based on English. Other works were performed in Chinese (Liu et al., 2022), Turkish (Şura Genç and Surer, 2023; Geçkil et al., 2018) and Spanish (Oliva et al., 2022; García-Ferrero and Altuna, 2024). To the best of our knowledge, this is the first work that fully addresses the study of clickbait detection and spoiling in the Italian language. Moreover, we propose a novel task, i.e., *clickbait neutralisation*, which aims at filling the curiosity gap by rewriting the headline leveraging the information of the spoiler.

7.3 Dataset

7.3.1 Dataset Creation

Data were collected from fourteen news websites[2], notorious for acting as news aggregators, engaging in plagiarism, lacking fact-checking, and using sensational headlines to draw in readers. In all the websites, articles are labelled according to specific categories; we decided to focus on four macro-categories: *health*, *science*, *economy*, and *environment*. These categories have been selected to cover some of the most frequent - and potentially hazardous - domains where clickbait is usually found. Since the categories varied a lot from website to website, we manually mapped each category into one of the four macro categories under analysis. Two annotators, knowledgeable in the area, were then provided with the headlines and the related articles and were asked to label whether a headline was clickbait. For aiding in this task, we have used as reference the clickbait measure as computed by Capozzi Lupi et al. (2023). Eventually, given the clickbait dataset, the two annotators were required to extract the gold spoilers from the article's text and to produce the neutralised forms for each headline. To this end, we employed an author-reviewer strategy (Tekiroğlu et al., 2020): an LLM (ChatGPT `gpt-3.5-turbo-0125`[3]) was used to generate both the spoilers and the neutralised forms (author component)[4], and the native Italian speaking annotators were asked to manually post-edited the generations (reviewer component).[5] This procedure was proven to be more effective and less time-consuming than writing the data from scratch (Bonaldi et al., 2022). Similarly to Chapter 3, to assess the amount of post-editing required, we employed *Human-targeted Translation Edit Rate* (HTER; Snover et al., 2006), which quantifies the minimum number of editing operations needed between a machine-generated text and its post-edited counterpart. Further details on the metric can be found in Section 4.3.5.

The obtained HTER results for the spoiler generation (*0.4*) are higher than those computed upon the neutralisation (*0.3*), in par or slightly lower than the 0.4 threshold. The high HTER values, especially for the spoiler annotation, can be attributed

[2]Essere Informati, TGNewsItalia, Voxnews, DirettaNews, Informati, Italia, Jeda News, News Cronaca, TG5Stelle, TG24-ore, ByoBlu, Mag24, WorldNotix, lo sapevi che, Fortementein

[3]https://chat.openai.com

[4]In Appendix D.1.3 we provide the prompt employed

[5]Details in Appendix D.1.2

Category	Headline	Article	Clickbait	Spoiler	Neutralised title
Health	Frutto o fiore? gustosissima e attraente, una celebrità sulle nostre tavole, sveliamo chi è	Tutti la conosciamo, immancabile sulle nostre tavole, celebre in tutto il mondo ma misteriosa la sua natura, frutto da gustare o fiore...	True	La fragola	Fragola: gustosissima e attraente, una celebrità sulle nostre tavole
Science	Scoperto un metallo che si auto-ripara. Scienziati sbalorditi	Il recente esperimento ha rivelato un fenomeno straordinario...	True	Il platino	Il metallo che si auto-ripara: il platino
Health	Una malattia che colpisce 500mila persone	Parliamo di una malattia sistemica cronica mediata dal sistema immunitario che interessa...	True	La psoriasi colpisce circa 500 mila persone	La psoriasi: una malattia che colpisce circa 500mila persone in Italia
Environment	Zanzare, ecco come eliminarle senza insetticidi	Con l'arrivo del caldo, anche le zanzare si fanno largo nelle nostre case o nei nostri giardini...	True	Per eliminare una volta per tutte le zanzare dalla vostra casa, dovreste acquistare un pipistrello	Zanzare, ecco come eliminarle senza insetticidi: basta acquistare un pipistrello

Table 7.1: An excerpt of the presented dataset showing the most relevant fields. Article bodies are shortened for space reasons. Translated text can be found in Table D.3 (Appendix D.2).

to the model's tendency to generate spoilers comprising more details than those necessary to fill the curiosity gap. While in some cases a simple deletion was sufficient, in others the annotator had to rewrite the spoiler almost completely. Regarding the annotation of the neutralisation texts, the higher results are a consequence of the spoiler generation, as the model was required to generate them simultaneously.

With this, we have generated the *gold* set of the dataset, in which all the instances were manually annotated. Further details regarding the dataset creation can be found in Appendix D.1. To expand this set, we have used a clickbait classifier (see Sect. 7.4.1) to automatically detect clickbait headlines. This new set of data, automatically annotated, constitutes the *silver* set of our dataset. Several examples of dataset entries are provided in Table 7.1.

7.3.2 Dataset Analysis

The complete *ClickBaIT* dataset consists of 4,144 entries. Each entry includes the following fields: (i) **source website**, that specifies the source of the article; (ii) **publication date**, which is captured from the original source; (iii) **headline text**; (iv) **article text**; (v) **original URL**; (vi) **macro category** inferred from the

Set	Clickbait (%)	Non-clickbait (%)	Total
gold	698 (53%)	629 (47%)	1,327
Silver	1,563 (56%)	1,224 (44%)	2,787
Total	2,261	1,853	4,114

Table 7.2: Size of the presented dataset, considering both gold and silver sets.

original category extracted from the source; (vii) **image URL** associated with the article as specified in the source; (viii) **clickbait annotation**; (ix) the **associated spoiler**; and (x) the **neutralised version** of the title.

Table 7.2 shows the main statistics of the final version of the dataset. The gold set is manually annotated and thus contains high-quality information. Additionally, the silver set has been annotated automatically as described and therefore contains a larger number of instances.

To gain a deeper understanding of the content of the dataset, we have used Variationist (Ramponi et al., 2024), a tool that allows us to inspect useful statistics and patterns in textual data. Upon inspection of the data, we have detected several patterns frequently used for generating the curiosity gap.

Of course, one of the most common strategies used in clickbait headlines is the formulation of a question that is later answered in the article, even though sometimes it is not. In the instance "*Quanto è green il gas?*" (*How green is gas?*), the article explains that gas is not considered green. Another frequent strategy we have detected is the introduction to the content of the article, which invites the reader to click it: *Beve un cucchiaio di aceto di mele nell'acqua tutti i giorni, ecco cosa succede* (*Drinks a tablespoon of apple cider vinegar in water every day, this is what happens*).

Another usual pattern is the reference to enumerations, frequently using round and manageable numbers such as 10, 8, and 5. This can be done for introducing numbered content, as in "*Le 10 fantasie femminili più segrete*" (*The 10 most secret female fantasies*), or even to generate a reaction in the reader: "*Hai solo 10 secondi per salvarti. Ecco cosa devi fare:*" (*You only have 10 seconds to save yourself. Here's what you have to do:*). Other means can be used to make headlines noticeable, such as introducing text in all caps, using striking vocabulary or even punctuation marks, as in "*[ALLARME] Truffa AUTO USATE, fate attenzione!*" (*[ALERT] USED CAR scam, beware!*).

See Table D.2 (Appendix D.1.2) for a collection of patterns that have been considered during the manual annotation of the dataset. Besides, Appendix D.2 includes a graphical summary of the dataset, while its interactive version can be accessed online.[6] Details are provided in Appendix D.3.

7.4 Experimental Design

Our experimental design comprises three steps: *clickbait detection*, *spoiler generation*, and *clickbait neutralisation* (see Figure 7.1).

[6] https://oaraque.github.io/ClickBaIT/clickbait.html

7.4.1 Clickbait Detection

This is the first and most basic task aimed at addressing the clickbait phenomenon. To explore the effect of using additional data in the training process, we use the Webis-Clickbait-17 (Potthast et al., 2018a), an English dataset containing clickbait that is also annotated in a binary fashion.

Following the insights by Araque et al. (2023), we use the training on English data to improve the classification of Italian data. The main idea is to harness the availability of large amounts of English data, generating a compound dataset with a lower amount of Italian instances. To do so, a multilingual mixture dataset is created so that 35% of the final dataset comprises Italian instances, while the rest are in English.

We model the detection challenge as a binary classification task: clickbait/non-clickbait. To study the complexity of the task, we explore two different models for classification: (i) a DistilBERT (Sanh et al., 2019) (`distil-base-multilingual-cased`[7]) model trained in a multilingual setting, and (ii) the Llama3-8B language model (`meta-llama/Meta-Llama-3-8B`[8]). The composed dataset has been split into train and test splits, which have been used to fine-tune and evaluate these models, respectively.

To assess the effect of using a mixture of both English and Italian instances in the dataset, we evaluate the performance of the two models in a monolingual setting (e.g., fine-tuning in Italian and predicting in the same language) as well as the multilingual variant (e.g., fine-tuning in English and Italian text, and predicting on Italian instances).

7.4.2 Spoiler Generation

The spoiler generation task consists in generating a short message that fulfils the curiosity gap present in a given clickbait title, by extracting the information from the linked article. To this end, we tested `LLaMAntino-3-ANITA-8B-Inst-DPO-ITA` (LLaMAntino-3-8B hereafter) (Polignano et al., 2024) on our clickbait dataset. The model was tested both in in-context learning (zero- and few-shot) and fine-tuning settings.

Building on prior research that frames spoiler generation as a Question Answering task (Woźny and Lango, 2024), we prompt the model to rewrite clickbait headlines as questions and extract the corresponding answers, i.e., the spoilers, from the linked articles.

7.4.3 Clickbait Neutralisation

The best-performing configuration was employed for the neutralisation of the clickbait headlines. To this end, we instructed the LLM to perform a style transfer task, from a clickbait headline style to a more journalistic one, while integrating the spoiler information into the original headline.

	zero-shot			few-shot			fine-tuning		
	R1	RL	SemSim	R1	RL	SemSim	R1	RL	SemSim
Headlines	0.189	0.157	0.567	0.250	0.221	0.667	0.260	0.234	0.659
Questions	0.271	0.249	0.645	0.286	0.258	0.630	0.250	0.224	0.646

Table 7.3: `LLaMAntino-3-8B` results for the spoiler generation task. We report ROUGE 1 and L (R1, RL) and semantic similarity (SemSim).

7.5 Results and Discussion

7.5.1 Evaluation Metrics

Firstly, for the evaluation of the clickbait detection task, we use macro-averaged precision, recall, and F-score, which allow us to assess performance reliably even in unbalanced scenarios. For the generation tasks, we assessed lexical similarity through ROUGE score (Lin, 2004) and semantic similarity; for the latter, text embeddings computed using `sentence-bert-base-italian-xxl-uncased`[9] were compared using cosine similarity. A detailed description of all metrics employed can be found in Section 2.4.5 of Chapter 2.

7.5.2 Clickbait detection

Table 7.4 shows the results of the evaluation in the task of clickbait classification. As expected, introducing data instances in English improves the performance in Italian. In the case of classification in Italian, we see a staggering improvement for the Llama3 model of 8.43 points. We argue that augmenting the training set with instances in a diverse language is an effective strategy that may generalise to other tasks. In particular, when annotated data in a target language is scarce, supplementing the training set with semantically similar instances from another language represents a viable solution to improve performance, as confirmed by the gains observed here for Italian clickbait detection and in line with prior findings in low-resource settings (Araque et al., 2023).

We also see that the best model for the classification of clickbait is the one obtained with Llama3, trained with both English and Italian data. Hence, we use this model to predict on the silver set of our dataset.

7.5.3 Spoiler Generation Results

Results for the spoiler generation task are reported in Table 7.3. We evaluated the capabilities of `LLaMAntino-3-8B` in both in-context learning scenarios (zero- and few-shot) and through fine-tuning. As inputs, we used clickbait headlines and questions generated by ChatGPT, instructing the model to execute a Question Answering task for the latter. When using headlines as input, few-shot and fine-tuning approaches outperform zero-shot methods. Few-shot approaches demonstrate higher performance in terms of semantic similarity, while fine-tuning exhibits stronger lex-

[7] `https://huggingface.co/distilbert-base-multilingual-cased`
[8] `https://huggingface.co/meta-llama/Meta-Llama-3-8B`
[9] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

Test	Train	Model	Prec.	Rec.	M-F1
EN	EN	DistilBERT	67.15	70.34	66.94
		Llama3	68.42	66.46	67.18
	EN+IT	DistilBERT	70.28	70.14	70.12
		Llama3	**71.20**	**71.15**	**71.15**
IT	IT	DistilBERT	68,85	70.47	68.65
		Llama3	66.96	67.19	67.07
	EN+IT	DistilBERT	72.87	74.85	71.77
		Llama3	**76.32**	**75.51**	**75.50**

Table 7.4: Results for Clickbait detection. The 'Test' and 'Train' columns indicate the languages of the test and train sets, respectively.

ical adherence to the source document, as reflected in ROUGE scores. This can be attributed to the few examples provided in the few-shot approach, which make the model aware of the task while allowing more creative outputs (resulting in lower ROUGE scores). Conversely, the fine-tuned model learned from the training data to adhere more closely to the source article, which comes at the expense of producing semantically richer responses (evidenced by lower SemSim scores).

Interestingly, casting spoiler generation as a question-answering task yields higher results in the zero-shot setting compared to using headlines as input. However, the results for few-shot and fine-tuning scenarios tend to be on par. This can be explained by the fact that headlines may contain multiple gaps that the human-annotated dataset accounted for, but the non-supervised "question generation" module could not fully capture. Generally, this approach leads to sufficiently good results; however, we believe that more attention should be given to the quality of the questions, either through more efficient prompts or with human-generated/curated data.

7.5.4 Clickbait Neutralisation Results

In Table 7.6, we report the results for clickbait neutralisation. For this task, we prompted `LLaMAntino-3-8B` with a few-shot approach, employing the spoilers generated with the three configurations of the previous experiments (headlines as input). Using spoilers generated with the fine-tuned models leads to higher results both for lexical and semantic metrics. Interestingly, scores tend to increase when the training complexity of the input data increases. In Table 7.5, we report examples of headlines along with their generated spoilers (through the fine-tuned model) and their neutralisation.

Headline	Spoiler	Neutralisation
“Juventus in Serie B”: perché c’è panico tra i tifosi, la scoperta delle ultime ore	15 punti di penalizzazione	Juventus in grave difficoltà: 15 punti di penalizzazione e il rischio di cadere in Serie B
Lutto tremendo nello sport italiano, morto giovanissimo dopo un malore	“Samuel Dilas era un giocatore di pallacanestro che militava nel Virtus Lumezzane a Brescia, in Serie B” e “aveva 24 anni” e “era alto 206 centimetri” e “nato a Novellara (Reggio Emilia)” e “aveva un padre di nome Torsen, una madre di nome Chiara e una sorella minore di nome Maia” e “era in convalescenza dopo una polmonite” e “era arrivato alla Virtus Lumezzane nella scorsa stagione”.	Tragico decesso del pallacanestrista Samuel Dilas, 24 anni, ex convalescente da polmonite e giocatore della Virtus Lumezzane
Un papà si rifiuta di mangiare accanto a un bambino Down di 5 anni, il cameriere decide di fare questo	Il cameriere ha sbattuto fuori il maleducato padre che voleva essere spostato a causa della presenza di un bambino con sindrome di Down.	Un cameriere espelle un cliente maleducato che chiede di essere spostato per non sedersi accanto a un bambino con sindrome di Down.
È doloroso e si forma tra le dita dei piedi, ecco come rimuoverlo	“L’occhio di pernice è causato principalmente dalla pressione della scarpa che favorisce la formazione di un’ispessimento di pelle che provoca dolore, in quanto è soggetto all’attrito tra le dita. Per rimuovere l’occhio di pernice è fondamentale ammorbidire prima la zona interessata per poi provare a rimuovere l’ispessimento utilizzando rimedi naturali senza dolore e in modo semplice."	Come rimuovere l’occhio di pernice, un problema di pressione e attrito causato dalle scarpe
La chiamano “LA BOMBA” la miscela che in sole 24-48 ore elimina influenza, raffreddore e tosse	Lo zenzero è un rimedio naturale per il trattamento di tosse, raffreddore e influenza. La miscela limone, zenzero e miele è ideale per alleviare i sintomi delle comuni malattie. Basta prendere 2 o 3 cucchiai della miscela naturale, riempire una tazza con acqua calda e lasciare in infusione per 3 o 4 minuti.	Miscela naturale di limone, zenzero e miele allevia i sintomi di tosse, raffreddore e influenza in pochi giorni.

Table 7.5: Examples of clickbait headlines, along with the automatically generated spoiler and neutralised version.

input data	R1	RL	SemSim
zero-shot	0.250	0.212	0.675
few-shot	0.265	0.223	0.706
fine-tuning	**0.286**	**0.247**	**0.715**

Table 7.6: Neutralisation generation results. Automatically generated spoilers from the previous experiments were used as input for the few-shot generation of the data. We report ROUGE 1 and L (R1 and RL) and the semantic similarity scores.

7.6 Conclusion

This work presents *ClickBaIT*, a novel Italian dataset for clickbait modelling, as well as a diverse set of experiments to assess the effectiveness of current models for clickbait detection, spoiling and neutralisation. The dataset includes news articles that have been manually annotated to indicate the presence of clickbait, spoilers associated with clickbait headlines, and their respective neutral headlines.

The experiments explore the effectiveness of current NLP methods for the modelling of clickbait headlines in Italian through *ClickBaIT*. The evaluation for clickbait detection shows how training data can be augmented in a multilingual setting, which leads to classification improvements that are in line with previous research (Araque et al., 2023). The generation experiments, for both spoiling and neutralisation, evidence that the evaluated model does benefit from in-domain knowledge extracted from the proposed dataset. As seen, these informed generations are more accurate and align better with the gold text.

Considering the effect of clickbait, we argue that while linked articles might be harmless, the lack of accuracy can have a detrimental effect on readers. This is clear when considering certain sensitive domains such as health. Thus, we hope that this work facilitates future research on the topic, for example, by addressing the link between clickbait and misinformation, considering both in a unified framework.

8

Conclusions

The proliferation of misinformation in the digital age constitutes one of the most pressing challenges of the Post-Truth era, threatening democratic discourse, public health, and institutional trust. Automated fact-checking has emerged as a necessary complement to the activity of human fact-checkers, whose capacity to keep pace with the scale and velocity of online misinformation is inherently limited. Within this landscape, knowledge-driven generative approaches have demonstrated considerable promise for producing accurate, readable verdicts that explain the veracity of misleading claims. Yet, prior work has largely operated under simplified conditions that rarely hold in practice: claims are assumed to be well-formed and expressed in standard journalistic language; reliable knowledge is assumed to be readily available and pre-paired with the claim; and the communicative contexts in which misinformation actually circulates are largely ignored. These assumptions constrain the practical applicability of existing systems and widen the gap between academic research and real-world deployment.

This thesis has addressed these limitations by systematically investigating the following overarching question: *How can knowledge-driven generative approaches for countering misinformation be employed to address increasingly realistic scenarios?* To answer it, we decomposed the problem into five interconnected research questions, each progressively relaxing a distinct idealised assumption and advancing the state of the art along four critical dimensions: (i) adapting to the communication styles of the platforms where misinformation circulates; (ii) expanding geographical and linguistic coverage beyond English-centric approaches; (iii) operating effectively when high-quality, paired knowledge is unavailable; and (iv) handling forms of misinformation that extend beyond discrete, verifiable claims. What follows traces the contributions of this thesis through the lens of these research questions, highlighting both what was achieved and how each step exposed the limitations that motivated the next.

The starting point was foundational. **RQ1** asked how language models can leverage reliable knowledge to generate responses to misleading claims, and in par-

ticular: *how salient information can be extracted from fact-checking articles to support appropriate verdict generation* (RQ1.1), and *whether stylistic adaptation to the conventions of different fact-checking organisations can be achieved through multitask learning without sacrificing factual quality* (RQ1.2). Chapter 3 addressed both sub-questions by benchmarking a comprehensive range of summarisation-based approaches across two datasets reflecting distinct journalistic styles. The hybrid pipeline, combining extractive and abstractive summarisation, proved consistently competitive, confirming that structured knowledge extraction is a valuable step prior to generation. The multitask learning experiments further demonstrated that stylistic adaptation across organisations is feasible, providing an initial affirmative answer to both RQ1.1 and RQ1.2.

However, the style dimension explored in Chapter 3 was confined to variations between professional fact-checking outlets. In practice, a substantial share of misinformation spreads on social media platforms, where the communicative register differs markedly from journalistic prose: claims are informal, emotionally charged, and frequently accompanied by personal commentary. Responding to such claims in a dry, journalistic register risks undermining the effectiveness of the counter-message. **RQ2** therefore asked whether generated verdicts can adapt to the interlocutor's communication style while maintaining factual accuracy and incorporating empathetic framing. Chapter 4 addressed this question by introducing `VerMouth`, a large-scale, general-domain dataset for verdict generation in social media contexts, constructed through a human-in-the-loop pipeline combining instruction-based LLMs with human post-editing. Models trained on `VerMouth` consistently outperformed those trained on journalistic data when evaluated on social media-style claims, and both automatic and human evaluation confirmed that style- and emotion-adapted verdicts are widely preferred for this type of input. The chapter thus demonstrated that the communicative dimension of automated fact-checking is not peripheral to the task but central to its effectiveness.

This stylistic flexibility proved to be more than an academic refinement: engagement with fact-checking organisations, through both qualitative interviews and evaluation within EU-funded projects, confirmed that the ability to adapt automatically generated drafts to an organisation's communicative style is a practical necessity rather than a stylistic preference, since a verdict requiring extensive post-editing to match house style can erode the time savings that automation is meant to provide. The same engagement confirmed the core premise motivating this line of work: automating verdict drafting measurably eases the workload of fact-checkers, provided the output remains editable rather than treated as a final product.

Yet both Chapter 3 and Chapter 4 shared a critical limitation: all resources and experiments were confined to English. Misinformation, however, is a global phenomenon, and the same misleading claim may circulate simultaneously across multiple linguistic communities. Fact-checking organisations operate in dozens of languages, and the knowledge they produce is not always available in the language of the claim to be addressed. **RQ3** asked whether knowledge-driven approaches can generate reliable verdicts in multilingual scenarios, including the challenging cross-lingual case in which supporting evidence is expressed in a language different from the claim. Specifically, it investigated *whether multilingual language models*

can counter misinformation when reliable knowledge is available in the same language as the claim (RQ3.1), and *whether they can do so when available knowledge exists only in a different language* (RQ3.2). Chapter 5 tackled both sub-questions by introducing **EuroVerdict**, a multilingual dataset spanning eight European languages developed in collaboration with professional fact-checkers, and by shifting from the summarisation-based paradigm of earlier chapters to an instruction-based LLM evaluated under zero-shot, Chain-of-Thought, and fine-tuning configurations. Fine-tuning proved to be the dominant strategy across all languages, and the performance degradation observed when evidence was provided in a different language from the claim remained modest, demonstrating that multilingual LLMs are sufficiently robust for cross-lingual verdict generation and yielding positive answers to both RQ3.1 and RQ3.2.

Across all three preceding chapters, however, a foundational assumption had remained intact: the existence of a relevant fact-checking article directly associated with the claim under examination. In real-world fact-checking workflows, this assumption fails with considerable regularity. Claims may concern events for which no fact-check has yet been published, or a relevant article may exist but not be paired with the claim at hand, requiring retrieval from a broader knowledge base. **RQ4** confronted this challenge directly by asking how accurate, faithful, and style-adapted verdicts can be generated when claim-specific knowledge is unavailable, decomposing the problem into three sub-questions: *how to generate high-quality verdicts when a fact-checking article exists but is not directly paired with the claim* (RQ4.1); *how to do so when no claim-specific fact-checking article exists and the system must rely on general reliable evidence* (RQ4.2); and *how the communicative style of the claim affects overall Retrieval-Augmented Generation (RAG) pipeline performance* (RQ4.3). Chapter 6 addressed these through a comprehensive evaluation of RAG pipelines across a wide range of configurations, varying retrieval strategy, knowledge base composition, document granularity, claim pre-processing, and generation setup. LLM-based retrievers offered the greatest adaptability, while hybrid retrieval represented a computationally efficient alternative; fact extraction modules consistently improved retrieval performance, particularly for informal and emotionally charged claims, directly addressing RQ4.3. The chapter further introduced **First-AID**, a human-in-the-loop framework for constructing synthetic, knowledge-grounded dialogue corpora, addressing the upstream data scarcity that constrains fine-tuning in this setting. Taken together, these findings provided a positive answer to RQ4: RAG-based pipelines are viable for verdict generation under realistic knowledge conditions, even in the complete absence of pre-existing fact-checks. Two aspects of this evaluation, retrieval constrained to curated, trusted source lists rather than open web-scale retrieval, and model selection assessed across varying sizes, computational requirements, and languages rather than for raw performance alone, were later corroborated through subsequent engagement with fact-checking organisations, including both qualitative interviews and evaluation within EU-funded projects. This engagement confirmed source control and deployment-aware model selection as genuine operational priorities, rather than assumptions imposed solely by the research design. The outcome of this analysis helps inform the creation and deployment of automatic fact-checking systems that account for the different needs and constraints

of individual organisations, rather than assuming that a single model or configuration is suitable across all deployment contexts.

The preceding four research questions collectively assumed that misinformation presents itself as a discrete, verifiable claim that can be directly submitted to a knowledge-driven pipeline. This assumption, too, has its limits. Clickbait headlines exemplify a form of misinformation in which the deception does not reside in a falsifiable statement but in the gap between what a headline promises and what the linked article delivers. Before such content can be addressed by claim-based fact-checking systems, the headline must first be completed with the factual information present in the source article, a process we refer to as *factuality completion*. **RQ5** asked how knowledge-driven approaches can address this more complex form of misinformation. Chapter 7 tackled this question by introducing **ClickBaIT**, a novel Italian dataset for clickbait detection, spoiler generation, and the new task of clickbait neutralisation, which reformulates sensationalist headlines into factually complete, journalistically sound alternatives by integrating information extracted from the source article. The evaluation confirmed that fine-tuned LLMs can perform this task with meaningful accuracy, and that framing spoiler generation as a question-answering task improves downstream neutralisation quality. Beyond the immediate results, this chapter extended the scope of knowledge-driven generation to a form of misinformation that lies outside the reach of conventional fact-checking frameworks, opening a new research avenue for the Italian language and, more broadly, for non-English misinformation contexts.

Considered as a whole, the five research questions and their corresponding chapters trace a coherent trajectory from controlled, English-language, resource-rich settings toward increasingly realistic, multilingual, and knowledge-scarce scenarios. At each step, the limitations of the preceding approach shaped the direction of the subsequent contribution, resulting in a body of research that advances not only individual components of automated fact-checking but the field's broader capacity to operate under the complex conditions of real-world misinformation ecosystems.

Despite the contributions presented in this thesis, several limitations remain. The following discusses the most significant ones, alongside the future research directions they motivate.

Faithfulness and Knowledge Reliability

Across all chapters, conditioning the generative model on external evidence, whether fact-checking articles, retrieved documents, or source articles, consistently helped reduce factual inconsistencies relative to unconstrained generation. Nevertheless, this conditioning does not eliminate the problem. Verdicts can drift from the supporting evidence, particularly when the retrieved context is noisy, incomplete, or only partially relevant to the claim at hand, as observed in the RAG experiments of Chapter 6. Closely related to this is the question of knowledge reliability. Throughout the thesis, retrieved knowledge is treated as trustworthy by assumption. This is a reasonable approximation when the knowledge base consists of curated fact-checking articles, but it becomes increasingly difficult to defend as the system moves toward web-scale or heterogeneous sources. Source credibility is not modelled at any point in the retrieval pipelines presented, meaning that the quality of the generated verdict

is ultimately bounded by the quality of the evidence fed to the generator. Future work should address both dimensions jointly, developing faithfulness-aware generation strategies and fine-grained attribution mechanisms on the generation side, while integrating source credibility scoring and evidence triangulation across multiple independent sources on the retrieval side.

Collaborating fact-checkers further pointed to a related, broader imbalance in this thesis: the bulk of the technical effort addressed the generative side of the pipeline, while evidence retrieval and source reliability received comparatively less attention, despite being equally critical to deployment, since a fluent and well-structured verdict is of limited value if the underlying evidence cannot be trusted. A related and currently unaddressed requirement is generation-time transparency: none of the systems presented here expose to the user which specific document or passage a given statement in the verdict was drawn from. Such fine-grained attribution, linking generated content back to the retrieved chunk or source document, is a precondition for a fact-checker to verify and trust a system's output quickly enough for it to be usable in an operational, time-constrained setting. More broadly, advancing retrieval reliability and generation-time transparency stand out as two research directions of particular importance for the field, as progress on these fronts is a necessary condition for moving knowledge-driven verdict generation closer to direct applicability in real-world fact-checking workflows.

Evaluation of Generated Verdicts

Assessing the quality of the outputs produced by knowledge-driven generative systems is an inherently difficult problem. Verdicts are complex communicative acts that must simultaneously satisfy factual, stylistic, and pragmatic criteria, yet no established evaluation framework captures all of these dimensions adequately. The evaluation conducted across the chapters of this thesis reflects this difficulty: automatic metrics provide partial and sometimes conflicting signals, human evaluation is more informative but costly and difficult to scale, and the results produced by different approaches do not always converge. This is not a limitation specific to this thesis but a systemic gap in the field. Future work should prioritise the development of comprehensive, task-specific evaluation frameworks for verdict quality, as well as shared benchmarks and annotation guidelines that account for the multidimensional nature of the task across languages and communicative contexts.

Adaptation to User Characteristics

The style adaptation pursued in this thesis operates at the level of platform register and emotional tone: verdicts are adapted to match the informal language of social media or to respond empathetically to emotionally charged claims. The individual receiving the verdict, however, is not modelled. The effectiveness of a correction is known to depend heavily on who receives it, since prior beliefs, political orientation, cultural background, and susceptibility to specific argument types all condition how a counter-message is processed and accepted. A verdict that is factually accurate and stylistically appropriate may still fail to persuade if it does not account for these dimensions. Future work should incorporate richer user modelling into verdict generation, moving beyond surface-level stylistic adaptation toward systems that can tailor their argumentative choices to the characteristics of the target interlocutor.

This line of research would benefit from closer engagement with the psychological literature on belief revision, inoculation theory, and narrative persuasion, which provides well-established models of how individuals respond to corrections depending on factors such as political polarisation and cultural context.

Scope of Misinformation Types

The work presented in this thesis focuses on two manifestations of misinformation: discrete, verifiable claims that can be addressed through knowledge-driven verdict generation, and clickbait headlines that require factuality completion before entering a fact-checking pipeline. Numerous other forms of misinformation that are prevalent and potentially harmful online lie outside this scope, including multimodal misinformation arising from inconsistencies between images and accompanying text, coordinated inauthentic behaviour that operates at the level of networks rather than individual claims, and satirical content that is misread or deliberately reframed as factual. Future work should investigate how the knowledge-driven generative framework developed in this thesis can be extended to these settings. The factuality completion paradigm introduced in Chapter 7 for clickbait, in particular, represents a promising conceptual bridge toward other forms of complex misinformation where the deceptive element cannot be directly submitted to a claim-based fact-checking pipeline.

Misinformation is not a new phenomenon, but the Post-Truth era has given it unprecedented reach and legitimacy, eroding the shared epistemic foundations on which democratic societies depend. Timothy Snyder's dictum that *"to abandon facts is to abandon freedom"* (Snyder, 2017) captures precisely what is at stake: not merely the accuracy of individual claims, but the conditions that make free, informed collective life possible. This thesis has been written in the conviction that the technical and the societal dimensions of this challenge are inseparable: building systems that are more faithful, more linguistically inclusive, and more aware of the humans they are designed to serve is not merely a research objective but a contribution to a broader collective effort to restore trust in information.

Meeting this challenge, however, requires more than technical progress alone. The work presented in this thesis sits at the intersection of natural language processing, cognitive psychology, and the social sciences, and the connections between these disciplines are not incidental but constitutive. Understanding how people process corrections, how beliefs resist revision, and how communicative context shapes the reception of factual information are not peripheral concerns but central ones for anyone building systems intended to operate in the real world. Sustaining and deepening this interdisciplinary dialogue is, in this view, as important as advancing individual system components.

This thesis does not solve the problem of misinformation. The problem is too deeply rooted in the social, psychological, and political fabric of contemporary life for any single body of technical work to resolve. What it does is advance the state of the art toward automated systems that are more grounded in the conditions under which misinformation actually spreads: systems that can speak the language of the people they are meant to reach, reason over imperfect and incomplete knowledge,

and engage with the full complexity of deceptive content in the wild. The truth may no longer be self-evident in the Post-Truth era, but the effort to defend it remains both possible and necessary.

Bibliography

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Computation+ Journalism Symposium,(September)*.

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463. PMID: 36795592.

Amol Agrawal. 2016. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 268–272.

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. *ArXiv*, cs.DB/1906.09198. Version 1.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Michelle Amazeen and Arunima Krishna. 2020. Correcting vaccine misinformation: Recognition and effects of source type on misinformation via perceived motivations and credibility. *SSRN Electronic Journal*.

Michelle A. Amazeen. 2015. Revisiting the epistemology of fact-checking. *Critical Review*, 27(1):1–22.

Michelle A Amazeen and Erik P Bucy. 2019. Conferring resistance to digital disinformation: The inoculating influence of procedural news knowledge. *Journal of Broadcasting & Electronic Media*, 63(3):415–432.

Cynthia Andrews, Elodie Fichet, Yuwei Ding, Emma S. Spiro, and Kate Starbird. 2016. Keeping up with the tweet-dashians: The impact of 'official' accounts on online rumoring. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 452–465, New York, NY, USA. Association for Computing Machinery.

Oscar Araque, María Felipa Ledesma Corniel, and Kyriaki Kalimeri. 2023. Towards a multilingual system for vaccine hesitancy using a data mixture approach. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 471–475, Venice, Italy. CEUR Workshop Proceedings.

Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv preprint arXiv:1808.05542*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020b. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine*

Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 215–236, Cham. Springer International Publishing.

Melisa Basol, Jon Roozenbeek, and Sander Van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, 3(1).

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

M. T. Bastos and J. Farkas. 2019. "Donald Trump is my President!" the internet research agency propaganda machine. *Social Media and Society*, 5(3):1–13.

Ian M Begg, Ann Anas, and Suzanne Farinacci. 1992. Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121(4):446–458.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, 21(11).

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Leticia Bode and Emily K. Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Leticia Bode and Emily K Vraga. 2018. See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33(9):1131–1140.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily Booth, Jooyoung Lee, Marian-Andrei Rizoiu, and Hany Farid. 2024. Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*, 60(2):440–457.

Kelly Born and Nell Edgington. 2017. Analysis of philanthropic opportunities to mitigate the disinformation/propaganda problem.

Nadia M Brashier and Elizabeth J Marsh. 2020. Judging truth. *Annual Review of Psychology*, 71:499–515.

Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. 2021. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118.

Joel Breakstone, Mark Smith, Sam Wineburg, Amie Rapaport, Jill Carle, Marshall Garland, and Anna Saavedra. 2021. Lateral reading: College students learn to critically evaluate internet sources in an online course. *Harvard Kennedy School Misinformation Review*, 2(1).

Adam S Brown and Larry A Nix. 1996. Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5):1088–1100.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Arthur Thomas Edward Capozzi Lupi, Alessandra Teresa Cignarella, Simona Frenda, Mirko Lai, Marco Antonio Stranisci, and Alessandra Urbinati. 2023. Debunker assistant: A support for detecting online misinformation. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 494–498, Venice, Italy. CEUR Workshop Proceedings.

Regina Cazzamatta. 2024. Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries. *New Media & Society*, 0(0):14614448241268896.

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Robert Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107:1753.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, 8(1):1–7.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095.

Michael D Cobb, Brendan Nyhan, and Jason Reifler. 2013. Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology*, 34(3):307–326.

S Cohen, C Li, J Yang, and C Yu. 2011. Computational journalism: A call to arms to database researchers, 148-151. In *5th Biennial Conference on Innovative Data Systems Research, CIDR*.

Josh Compton, Sander van der Linden, John Cook, and Melisa Basol. 2021. Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, 15(6):e12602.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646.

Nancy Dai, Wenting Yu, and Fei Shen. 2021. The effects of message order and debiasing information in misinformation correction. *International Journal of Communication*, 15:1039–1059.

Jonas De keersmaecker, David Dunning, Gordon Pennycook, David G Rand, Carmen Sanchez, Christian Unkelbach, and Arne Roets. 2020. Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, 46(2):204–215.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political Psychology*, 40:3–35.

Fred I. Dretske. 1981. *Knowledge and the Flow of Information*. MIT Press, Stanford, CA.

Ullrich Ecker, Jon Roozenbeek, Sander Van Der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. 2024. Misinformation poses a bigger threat to democracy than you might think. *Nature*, 630(8015):29–32.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Ullrich KH Ecker, Stephan Lewandowsky, Briony Swire, and Darren Chang. 2011. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3):570–578.

Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8):1087–1100.

Ullrich KH Ecker, Ziggy O'Reilly, Joshua S Reid, and Ee Pin Chang. 2020. The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, 111(1):36–54.

EFCSN. 2022. European code of standards for independent fact-checking organisations.

P Ekman and W V Friesen. 1971. Constants across cultures in the face and emotion. *J Pers Soc Psychol*, 17(2):124–129.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Don Fallis. 2014. The Varieties of Disinformation. In Luciano Floridi and Phyllis Illari, editors, *The Philosophy of Information Quality*, pages 135–161. Springer International Publishing, Cham.

Don Fallis. 2015. What Is Disinformation? *Library Trends*, 63(3):401–426. Publisher: Johns Hopkins University Press.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. *arXiv preprint arXiv:2011.05448*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Jennifer D Featherstone and Jingwen Zhang. 2020. Feeling angry: The effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9):692–702.

Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of the ACM*, 59(7):96–104.

James H. Fetzer. 2004a. Disinformation: The Use of False Information. *Minds and Machines*, 14(2):231–240.

James H. Fetzer. 2004b. Information: Does it Have To Be True? *Minds and Machines*, 14(2):223–229.

Luciano Floridi. 2005. Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2):351–370.

Christopher John Fox. 1983. *Information and Misinformation: An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming*. Greenwood Publishing Group.

Maik Fröbe, Benno Stein, Tim Gollub, Matthias Hagen, and Martin Potthast. 2023. SemEval-2023 task 5: Clickbait spoiling. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2275–2286, Toronto, Canada. Association for Computational Linguistics.

Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 87–95, New York, NY, USA. Association for Computing Machinery.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Iker García-Ferrero and Begoña Altuna. 2024. Noticia: A clickbait article summarization dataset in spanish. *arXiv preprint arXiv:2404.07611*.

Ayse Geçkil, Ahmet Anil Müngen, Esra Gündogan, and Mehmet Kaya. 2018. A clickbait detection method on news sites. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 932–937.

Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems*, 4(4):196–206.

Maria Glenski, Svitlana Volkova, and Srijan Kumar. 2020. *User Engagement with Digital Deception*, pages 39–61. Springer International Publishing, Cham.

Yigal Godler and Zvi Reich. 2017. Journalistic evidence: Cross-verification as a constituent of mediated knowledge. *Journalism*, 18(5):558–574.

Andrew Gordon, Ullrich KH Ecker, and Stephan Lewandowsky. 2019. Polarity and attitude effects in the continued-influence paradigm. *Journal of Memory and Language*, 108:104028.

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3):518–537.

Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking.

Lucas Graves and Federica Cherubini. 2016. The rise of fact-checking sites in europe. Technical report.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. 2022. Clickbait spoiling via question answering and passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7036, Dublin, Ireland. Association for Computational Linguistics.

Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In

Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Jayson Harsin. 2018. Post-truth and critical communication studies.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2698–2709, New York, NY, USA. Association for Computing Machinery.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2022. spacy: Industrial-strength natural language processing in python.

Liwei Hou, Po Hu, and Chao Bei. 2017. Abstractive document summarization via neural model with joint attention. In *National CCF conference on natural language processing and Chinese computing*, pages 329–338. Springer.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Edda Humprecht. 2019. Where 'fake news' flourishes: a comparison across four western democracies. *Information, Communication & Society*, 22(13):1973–1988.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Hollyn M Johnson and Colleen M Seifert. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6):1420–1436.

S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American Behavioral Scientist*, 65(2):371–388.

Anastasia Katsaounidou, Lazaros Vrysis, Rigas Kotsakis, Charalampos Dimoulas, and Andreas Veglis. 2019. Mathe the game: A serious game for education and training in news verification. *Education Sciences*, 9(2):155.

Sawinder Kaur, Parteek Kumar, and Ponnurangam Kumaraguru. 2020. Detecting clickbaits using two-phase hybrid cnn-lstm biterm model. *Expert Systems with Applications*, 151:113350.

Panayiota Kendeou, Reese Butterfuss, Jongseok Kim, and Megan Van Boekel. 2019. Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition*, 47(1):33–46.

Panayiota Kendeou, Emily K Walsh, Edward R Smith, and Edward J O'Brien. 2014. Knowledge revision processes in refutation texts. *Discourse Processes*, 51(5-6):374–397.

Md Lamiur Khan and Inayat Khan Idris. 2019. Recognise misinformation and verify before sharing: A reasoned action and information literacy perspective. *Behaviour & Information Technology*, 38(12):1194–1212.

Emma Jane Kirby. 2016. The city getting rich from fake news. *BBC News*, 5.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, page 16.

Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2023. Social media and financial news manipulation. *Review of Finance*, 27(4):1229–1268.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

James H Kuklinski, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. 2000. Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, pages 591–602.

Hiroto Kurita, Ikumi Ito, Hiroaki Funayama, Shota Sasaki, Shoji Moriya, Ye Mengyu, Kazuma Kokuta, Ryujin Hatakeyama, Shusaku Sone, and Kentaro Inui. 2023. TohokuNLP at SemEval-2023 task 5: Clickbait spoiling via simple Seq2Seq generation and ensembling. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1756–1762, Toronto, Canada. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Stephan Lewandowsky, John Cook, Ullrich KH Ecker, Dolores Albarracín, Michelle A Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn J Newman, Gordon Pennycook, Ethan Porter, et al. 2020. The debunking handbook 2020.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369.

Stephan Lewandowsky and Sander van der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384.

Justin Matthew Wren Lewis, Andy Williams, Robert Arthur Franklin, James Thomas, and Nicholas Alexander Mosdell. 2008. The quality and independence of british journalism. MediaWise.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai Kodur, Shruti Rijhwani, Minghan Ma, Peng Shi, Wen-tau Yih, Xilun Chen, Luke Zettlemoyer, and Scott Yih. 2023. How to train your DRAGON: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6164–6179, Singapore. Association for Computational Linguistics.

Jerry Liu. 2022. LlamaIndex.

Tong Liu, Ke Yu, Lu Wang, Xuanyu Zhang, Hao Zhou, and Xiaofei Wu. 2022. Clickbait detection on wechat: A deep model integrating semantic and syntactic information. *Knowledge-Based Systems*, 245:108605.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116:75–98.

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257.

Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5:337–348.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and predicting social correction on twitter. *arXiv preprint arXiv:2303.08889*.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oğuz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. *arXiv preprint arXiv:2101.00117*.

Pranav Malhotra, Kristina Scharp, and Lindsey Thomas. 2022. The meaning of misinformation and those who correct it: An extension of relational dialectics theory. *Journal of Social and Personal Relationships*, 39(5):1256–1276.

Elizabeth J Marsh, Andrew D Cantor, and Nadia M Brashier. 2016. Believing that humans swallow spiders in their sleep: False beliefs as side effects of the processes that support accurate knowledge. *Psychology of Learning and Motivation*, 64:93–132.

Cameron Martel, Gordon Pennycook, and David G Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, 5:1–20.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Stefano Menini, Daniel Russo, Alessio Palmero Aprosio, and Marco Guerini. 2025. First-AID: the first annotation interface for grounded dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 563–571, Vienna, Austria. Association for Computational Linguistics.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir D. Memon. 2020. The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic. *arXiv*, cs.SI/2011.05773. Version 2.

Karen J Mitchell and Marcia K Johnson. 2009. Source monitoring 15 years later: What have we learned from fmri about the neural mechanisms of source memory? *Psychological Bulletin*, 135(4):638–677.

N. Moratanch and S. Chitrakala. 2017. A survey on extractive text summarization. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6.

Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G Rand. 2021. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lena Nadarevic, Rolf Reber, Antje J Helmecke, and Duygu Köse. 2020. Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications*, 5(1):1–16.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland. Association for Computational Linguistics.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors. 2023. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada.

Christian Oliva, Ignacio Palacio-Marín, Luis F. Lago-Fernández, and David Arroyo. 2022. Rumor and clickbait detection by combining information divergence measures and deep learning techniques. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, ARES '22, New York, NY, USA. Association for Computing Machinery.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

James Pamment and Anneli Lindvall Kimber. 2021. Fact-checking and debunking: a best practice guide to dealing with disinformation.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Myrto Pantazi, Mikhail Kissine, and Olivier Klein. 2018. The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social Cognition*, 36(2):167–198.

Shivangi B Parikh and Pradeep K Atrey. 2018. Media-rich fake news detection: A survey. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 436–441.

Kari A Parker, Bobi Ivanov, and Josh Compton. 2012. Inoculation's efficacy with young adults' risky behaviors: Can inoculation confer cross-protection over related but untreated issues? *Health Communication*, 27(3):223–233.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Jian Peng, Raymond Kim-Kwang Choo, and Helen Ashman. 2016. Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 121–128.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.

Kara Pernice, Kathryn Whitenton, and Jakob Nielsen. 2019. *How People Read Online: The Eyetracking Evidence*, 2nd edition. Nielsen Norman Group.

Pythagoras N. Petratos. 2021. Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, 64(6):763–774. CIBER SPECIAL ISSUE: CYBERSECURITY IN CRISIS.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018a. Crowdsourcing a Large Corpus of Clickbait on Twitter. In *27th International Conference on Computational Linguistics (COLING 2018)*, pages 1498–1507. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018b. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 810–817. Springer.

Nicolas Pröllochs. 2022. Community-based fact-checking on twitter's birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 794–805.

Xiaoyan Qiu, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour*, 1(7):1–7.

Dragomir Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

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

Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2021. Bert, xlnet or roberta: The best transfer learning model to detect clickbaits. *IEEE Access*, 9:154704–154716.

Alan Ramponi, Camilla Casula, and Stefano Menini. 2024. Variationist: Exploring multifaceted variation and bias in written language data. *arXiv preprint arxiv:2406.17647*.

David N Rapp. 2016. The consequences of reading inaccurate information. *Current Directions in Psychological Science*, 25(4):281–285.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST.

Jon Roozenbeek and Sander Van Der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):1–10.

Jon Roozenbeek and Sander van der Linden. 2020. Breaking harmony square: A game that "inoculates" against political misinformation. *Harvard Kennedy School Misinformation Review*, 1(8).

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17.

Daniel Russo, Oscar Araque, and Marco Guerini. 2024. To click it or not to click it: An Italian dataset for neutralising clickbait headlines. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 829–841, Pisa, Italy. CEUR Workshop Proceedings.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023a. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2025a. Face the facts! evaluating RAG-based pipelines for professional fact-checking. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 846–865, Hanoi, Vietnam. Association for Computational Linguistics.

Daniel Russo, Fariba Sadeghi, Stefano Menini, and Marco Guerini. 2025b. EuroVerdict: A multilingual dataset for verdict generation against misinformation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16617–16634, Vienna, Austria. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2116–2129, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

DY Sakhare, Raj Kumar, and Sudiksha Janmeda. 2018. Development of embedded platform for sanskrit grammar-based document summarization. In *Speech and Language Processing for Human-Machine Communications*, pages 41–50. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Lawrence J. Sanna and Norbert Schwarz. 2006. Metacognitive experiences and human judgment: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science*, 15(4):172–176.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Kate Scott. 2021. You won't believe what's in this paper! clickbait, relevance and the curiosity gap. *Journal of Pragmatics*, 175:53–66.

Kate Scott. 2023. "Deceptive" clickbait headlines: Relevance, intentions, and lies. *Journal of Pragmatics*, 218:71–82.

Kate Scott and Rebecca Jackson. 2020. When everything stands out, nothing does. *Relevance theory, figuration, and continuity in pragmatics*, 8:167–192.

Sebastian Sequoiah-Grayson and Luciano Floridi. 2022. Semantic Conceptions of Information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2022 edition. Metaphysics Research Lab, Stanford University.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9(1).

Louis Shao, Stephan Gouws, Denny Rennie, Aliaksei Mustafa, Arun Tejasvi Chaganty, Aaron van den Oord, and Chris Dyer. 2017. Generating long and diverse responses with neural conversation models. *arXiv preprint arXiv:1701.03185*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on*

Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on twitter during the 2012 us presidential election: Rumor diffusion and correction. *New media & society*, 19(8):1214–1235.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation. *arXiv*, cs.CL/2010.13002. Version 2.

Andrew Shtulman and Joshua Valcarcel. 2012. Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2):209–215.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Timothy Snyder. 2017. *On Tyranny: Twenty Lessons from the Twentieth Century*. Tim Duggan Books, New York, NY.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Samanth Subramanian. 2017. Inside the macedonian fake-news complex. *Wired magazine*, 15.

Briony Swire-Thompson, John Cook, Lucy H. Butler, Jasmyne A. Sanderson, Stephan Lewandowsky, and Ullrich K. H. Ecker. 2021. Correction format has a limited role when debunking misinformation. *Cognitive Research: Principles and Implications*, 6(1):83.

Sille Obelitz Søe. 2021. A unified account of information, misinformation, and disinformation. *Synthese*, 198(6):5929–5949.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3):460–480.

Kjerstin Thorson, Emily Vraga, and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. volume 64, pages 131–148. Elsevier.

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

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Christian Unkelbach and Rainer Greifeneder. 2018. Experiential fluency and declarative advice jointly inform judgments of truth. *Journal of Experimental Social Psychology*, 79:78–86.

Christian Unkelbach, Alex Koch, Rui R Silva, and Teresa Garcia-Marques. 2019. Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, 28(3):247–253.

Christian Unkelbach and Sarah C Rom. 2017. A referential theory of the repetition-induced truth effect. *Cognition*, 160:110–126.

Joseph E. Uscinski and Ryden W. Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Joseph E Uscinski, Adam M Enders, Casey Klofstad, Michelle Seelig, John Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar Murthi. 2020. Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*, 1(3).

Toni GLA Van der Meer and Yan Jin. 2020. Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, 35(5):560–575.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Emily K Vraga and Leticia Bode. 2017. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645.

Emily K Vraga and Leticia Bode. 2020. Correction as a solution for health misinformation on social media. *American Journal of Public Health*, 110(S3):S278–S280.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

María Celeste Wagner and Pablo J. Boczkowski. 2019. The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital Journalism*, 7(7):870–885.

Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3):423–441.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *arXiv preprint arXiv:2401.00368*.

Shuai Wang, Xiang Zhao, Bo Li, Bin Ge, and Daquan Tang. 2017. Integrating extractive and abstractive models for long text summarization. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 305–312. IEEE.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Mateusz Woźny and Mateusz Lango. 2024. Generating clickbait spoilers with an ensemble of large language models. *arXiv preprint arXiv:2405.16284*.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. Dtca: Decision tree-based co-attention networks for explainable claim verification. *arXiv preprint arXiv:2004.13455*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, WWW '19, page 3600–3604, New York, NY, USA. Association for Computing Machinery.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Andrew P Yonelinas. 2002. The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3):441–517.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, 11(3).

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

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

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Şura Genç and Elif Surer. 2023. Clickbaittr: Dataset for clickbait detection from turkish news sites and social media with a comparative analysis via machine learning algorithms. *Journal of Information Science*, 49(2):480–499.

A

Chapter 3

A.1 Data Augmentation Details

A.1.1 Prompt and Model Selection

The first step when it comes to effectively leveraging LLMs for one's specific use case is prompt engineering. In our case, a carefully crafted prompt allows LLMs to produce quality claims and verdicts systematically, minimising the amount of post-editing required. Since ChatGPT's API was not yet available to the public when we began our experiments, the initial prompt testing was performed using GPT3.

Initial tests focused on finding the optimal prompt and parameters for our specific use case. The parameters we tested were the *temperature* T and the cumulative probability p for *nucleus sampling* (*Top-P*).

The T hyperparameter ranges between 0 and 1 and controls the amount of randomness used when sampling: a value of 0 corresponds to a deterministic output (i.e., picking the top-probability token exclusively from the vocabulary); conversely, a value of 1 provides maximum output diversity. Finding a balance between not being overly deterministic while remaining coherent was the goal when testing different temperature values - 0.7 and 1 were used when performing these initial tests.

The *Top-P* parameter also ranges between 0 and 1 and determines how much of the probability distribution of words is considered during generation. It was necessary to find a value that was not overly deterministic and that avoided using very rare words that may reduce coherence. During these initial tests, we tried using the default value of 0.5 as well as a Top-P value of 1, which includes all words in the probability distribution. We also tested using a Top-P of 0.9.

Determining the "optimal" prompt and parameters can be challenging because what makes a "good" personalised claim or verdict is subjective. Several factors were taken into account when selecting our prompts and parameters:

1. **Generalisability**: Do they perform well on a variety of claims, or do they only work well on specific ones (ie: it produces quality output for Covid-related claims, but struggle with claims about Brexit)?

2. **Variability**: Do the generated claims and verdicts vary between one another, or do they all follow similar patterns?

3. **Originality**: Do the generated claims and verdicts resemble the original too much? Do they contain the original claim or verdict verbatim?

4. **Coherence**: Do the generated claims and verdicts make sense? Are they coherent? Are they saying what the original claims and verdicts are, or do they instead say something unrelated?

5. **Amount of Post-Editing**: On average, how much post-editing is required for each of the generated claims and verdicts? What proportion of these claims and verdicts requires any post-editing at all?

Eventually, we opted for the following parameters: a temperature of 1 and a Top-P of 0.9. These parameters were used with OpenAI's Davinci model during initial tests. No changes were made when we switched to ChatGPT after its public API was released.

A.1.2 Post-Editing

The post-editing of the data and the prompt evaluation were carried out by two annotators, either native English speakers or fluent in English. The time needed for post-editing was heavily dependent on the type of data (claim versus verdict) and on the configuration (SMP-style versus emotional style). On average, the annotators were able to process 250 SMP-style claims and 200 emotional claims per hour [1]. For the emotional data, the time required to post-edit varied greatly depending on the emotion. Since verdicts are usually longer than claims and much more constrained, the post-editing process was much longer: on average, 150 SMP-style verdicts and 70 emotional verdicts were able to be post-edited per hour.

Not every claim and verdict required post-editing. Only ~19% of the generated SMP-style claims and ~68% of the emotional style claims were post-edited. Keep in mind that some of the generated emotional claims were discarded if a specific emotion and the content of the claim were mismatched, as this resulted in forced and unnatural combinations. Figure A.1 displays the distribution of post-edited, non-post-edited, and discarded claims. Fear and surprise were the emotions with the least amount of discarded data, but they also required the most post-editing.

As mentioned before, post-editing verdicts were a much longer process. Since verdicts are subject to stricter standards of quality (because they must be truthful and polite, for example) and are much longer on average, many more of them required post-editing. Figure A.2 shows this disparity: for some emotions such as

[1] The annotators were timed on several occasions during post-editing (always sessions of 20 minutes to make fair comparisons); the results reported in the paper are the average of the claim/hour amount measured in each session.

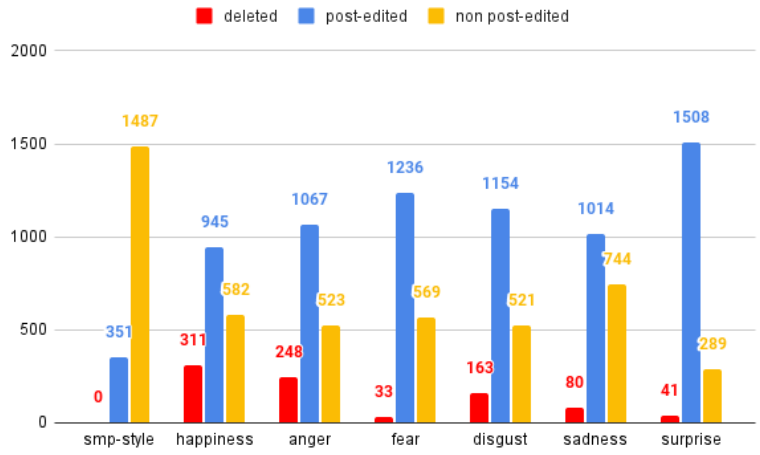


Figure A.1: Graph representing the number of ChatGPT-generated claims that were deleted, post-edited, or not post-edited

disgust and fear, there were fewer than 10 verdicts that did not require at least minimal post-editing. SMP-style verdicts also required post-editing at a much higher rate than SMP-style claims, although less than emotional style verdicts. In total, there were 1838 SMP-style verdicts and 2609 emotional style verdicts, resulting in post-editing rates of 91.4% and 96.3%, respectively.

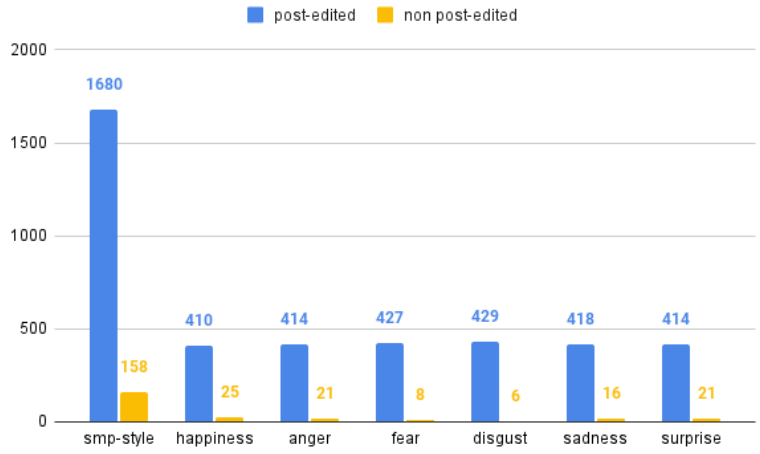


Figure A.2: Graph representing the number of ChatGPT-generated verdicts that were post-edited versus those that were not post-edited

A.1.3 Timing benefits of post-editing

We carried out an extra experiment to assess whether post-editing machine-generated data is more effective in terms of time than writing new data from scratch. To this end, we provided one of the annotators with 60 claims and asked them to write new tweets from scratch, 30 in SMP-style and 30 emotional tweets. In both cases, it took the annotator (expert in the field) around 23 minutes to create 30 new tweets (thus,

roughly 80 claims per hour as compared to the 250 SMP-style and 200 emotional tweets obtained with our pipeline). If in the creation of claims the time differences are considerable, we assume that this also applies to verdicts, which is a task that requires more constraints.

A.2 Detailed Guidelines

A.2.1 Claim Guidelines

1. **Generated claims copying original claims verbatim**: Sometimes, the wording of the original claim is copied verbatim in the generated claim. These should be rewritten to avoid resembling the original dataset. Determining whether a generated claim resembles the original claim "too much" can be subjective, so discretion must be used.
2. **Reoccurring Patterns**: Since the SMP-style claims have a lot more freedom to decide what sort of tone to adopt, they are much more diverse. With emotional style claims, the emotional component is an extra constraint that is applied during generation. This means that there are often recurring patterns which appear in the resulting generated claims: *"I'm livid!"*, *"'So sad to hear that X"*, *"Disgusting!"*, etc. If a pattern can be removed while preserving the overall emotional intent, then it is better to remove it entirely. Conversely, if removing a pattern also removes any *"emotion"* from the generated claim, then rewriting is preferred. In rare cases, the pattern can be kept, keeping in mind that too many occurrences may result in degraded performance during training.
3. **Hashtags**: Due to the prompt used, hashtags often appear in the generated claims. We noted two different phenomena that may occur and which require post-editing:

   (a) **Debunking hashtags**: There are some occurrences where a hashtag debunks a claim or works against the claim's intent. If a claim is about how vaccines are not effective, having the hashtag "#VaccinesSaveLives" is not appropriate. These hashtags can either be removed or edited to match the original intent.

   (b) **Unnecessary hashtags**: There are some hashtags which are so vague that they diminish the overall quality of the claim (such as *"#miracle"*, *"#goodjob"*, etc.). In emotional claims specifically, the emotion given in the prompt is turned into a hashtag (such as *"#happy"* or *"#sad"*). Any hashtags that fit these criteria are to be removed.

4. **Generated claims debunking original claims**: There are some instances where the generated claim actually *debunks* the original claim. These must be changed to reflect the intent behind the original claim. For example, if the original claim says that *"wearing masks causes dementia and hypoxia"* and the generated claim says *"False info alert: wearing a mask doesn't cause dementia and hypoxia"*, it should be rewritten to match the original claim.
5. **Dates and places**: The original claims contained many references to dates and places. These could be vague references (*"last year"*, *"in our nation"*, etc.) or

specific references (*"21 July 2021"*, *"in England"*, etc.). Any dates and places in the generated claims should match the original claim's level of specificity.

6. **Hallucinations**: Since claims do not necessarily need to be *true*, hallucinations can often be beneficial. Consider an example where the original claim says *"almost 300 people have died from the vaccine"*, but the generated claim contains a hallucination which states that *"291 people have died from the vaccine"* - this number is more specific, and this may give off the impression that the person knows what they are talking about.
7. **General formatting issues**: While rare, there are cases in which grammatical errors, typos, malformed hashtags (such as *"#endrape culture"*) or other formatting issues occur in generated claims. These should simply be corrected.

A.2.2 Verdict Guidelines

1. **Reoccurring Patterns**: As with generated claims, there are often patterns that occur often in generated verdicts. These should be removed or rewritten. Some examples of common patterns include *"thank you for X"*, *"I understand that you feel X"*, *"Let's continue to follow the recommended guidelines"*, etc.
2. **Calls to action**: As stated before, a *"call to action"* is a phrase or sentence which, as the name implies, calls upon the reader of the verdict to take action in some way. For example:

   "I understand your frustration, and while the proportion of BME students at Oxbridge has actually increased, I agree that more needs to be done to address the lack of diversity from disadvantaged areas. ***It's important that we continue examining the root causes of this inequality and work towards equal opportunities for all.*** *#diversitymatters #educationforall"*

   To avoid overtly political or polarising verdicts, many considerations need to be kept in mind when a call to action in a generated verdict is encountered.

   (a) **Is the call to action well-integrated into the verdict?** - If a call to action does not make a meaningful contribution to the overall quality of the verdict, then it will be removed. An example of a poorly-integrated call to action is *"let's focus on promoting peaceful and respectful discourse."* This call to action is broad and vague and should be removed.

   (b) **Is the call to action political or polarising?** - One must determine whether or not the call to action is actually political or polarising. With certain topics, it is simply impossible to avoid having a call to action that contains political elements (such as a claim about a politician, or a new law). We decided upon two possible approaches one can take when post-editing political calls to action.

   The *common sense approach* (or the *"reasonable person" approach*) is employed for a call to action expressing a political opinion on which most agree (such as *"demanding transparency and accountability from our government"*). This call to action can be kept.

   The *empathetic approach* is employed when dealing with opinions or thoughts that simply have no "correct" answers, but rely on one's own beliefs. It was

decided that the call to action should be changed to empathise with the claim writer's beliefs without agreeing or disagreeing with them.

For example, if a claim expresses pro-life opinions, and a call to action such as *"it's important to acknowledge the magnitude of lives affected by this issue"* exists, changing it to *"it's important to acknowledge the magnitude of lives affected by this issue* ***no matter what you believe****"* is empathetic, but also explicitly avoids picking one side or the other in a polarising situation like this.

(c) **How strong is a call to action, and who is the focus?** - Different actions require different amounts of effort. If a call to action asks for too much from someone, then it should be changed. Asking someone, for example, to *"advocate for stricter testing protocols"* may be asking too much of them, but asking them to *"hope that stricter testing protocols are implemented"* is not.

If a call to action is too strong, then it can either be *weakened* or *neutralised*. Weakening a call to action involves changing what is expected of the reader of the verdict: rather than *"we must personally take action to end child poverty immediately"*, one can post-edit the call to action to say *"let's try and do our part together to hopefully end child poverty one day"*.

Neutralising a call to action takes the focus off the reader entirely. This involves either putting the onus on someone else who may be more capable of solving the issue or not demanding action from anyone at all. Thus, *"****we*** *must continue to keep an eye out for potential side effects of the vaccines"* can be rewritten to *"****the experts*** *must continue to keep an eye out for potential side effects of the vaccines"*.

An example of not demanding action from anyone at all is: *"we must continue to advocate for those struggling to make ends meet"*. It can be post-edited as *"compassion and understanding for those struggling to make ends meet is crucial"*.

3. **Pronouns and Grammatical Personhood**: As the original verdicts sometimes contain first-person plural pronouns (*"we have contacted them for more clarification"*), and at other times contain first-person singular pronouns (*"I understand your frustration"*), there are inconsistencies regarding "who" is writing the verdict. The assumption one should take when post-editing is that each verdict is written by a single person. Therefore, if first-person plural pronouns are encountered, they should be changed to first-person singular pronouns.

   One exception exists: when the reader and writer of the verdict are grouped together, then first-person plural pronouns can be kept: *"surely* ***we*** *can all agree that this is a serious issue"*.

4. **Confirmations**: Sometimes a claim is fully or partially correct, and the original FullFact verdict notes this with a simple *"correct"*, or *"this is right, but..."*, but the generated verdict does not.

   In this case, adding a quick confirmation such as *"yes, you're right, but..."* or *"absolutely, it's a serious issue"* can be done as long as it does not reduce the

overall readability of the verdict.

5. **Missing information**: Sometimes the generated verdicts do not include information from the original verdicts. This can make the generated verdict easier to read without reducing its persuasiveness. If missing information negatively impacts how effective a verdict is, then it should be added.
6. **New claims**: Conversely, there are cases in which the generated verdicts actually include information that is not contained in the original verdict, but which is either objectively true or is a subjective statement. If the claims made in the generated verdict are provably false, then removing them is necessary. If they are provably true, or if they are a subjective statement or opinion which can not be concretely proven true or false, they can be kept or removed at the post-editor's discretion.
7. **General formatting issues**: As with generated claims, general formatting issues such as grammatical errors, typos, malformed hashtags, etc. should be corrected.

A.3 Experimental Details

A.3.1 Extractive Summarization Methods

Besides SBERT, we also considered other extractive summarization methodologies, i.e. **Lead-k**, which extracts the first k sentences from the article, and **LexRank** (Erkan and Radev, 2004), a graph-based unsupervised methodology which ranks the sentences of a document based on their importance by means of eigenvector centrality. We tested these approaches by extracting two sentence-long summaries from Fullfact articles. Subsequently, these summaries were evaluated against the gold verdicts. The results of this evaluation are shown in Table A.1. SBERT outperforms both Lead-2 and LexRank for all the metrics employed.

	R1	R2	RL	METEOR	BARTScore	BERTScore
Lead-2	0.21	0.06	0.15	0.26	-3.23	0.86
Lexrank	0.18	0.06	0.12	0.31	-2.89	0.86
SBERT	**0.25**	**0.09**	**0.18**	**0.33**	**-2.90**	**0.87**

Table A.1: Results for the extractive summarization methodologies on FF dataset.

A.3.2 Fine-Tuning Configuration

When fine-tuning, PEG_{base} was trained for 5 epochs with a batch size of 4 and a random seed set to 2022. To this end, we employed the HuggingFace Trainer [2] using the default hyperparameter settings, with the exception of the Learning Rate values and the optimisation method. Instead, we used the Adafactor stochastic optimisation method (Shazeer and Stern, 2018) and a Learning Rate value of 3e-05. The training was performed on a single Tesla V100 GPU, while the testing was

[2] https://huggingface.co/docs/transformers/main_classes/trainer

performed on a single Quadro RTX A5000 GPU. The checkpoint with minimum *evaluation loss* was employed for testing.

A.3.3 Decoding Configuration

At inference time, we employed *nucleus sampling* decoding strategy, setting the probability at 0.9, and repetition penalty at 2.0, for the verdict generation.

B

Chapter 4

B.1 Dataset Creation

B.1.1 Annotation Guidelines

Hereafter, we provide the annotation guidelines:

> Please open the dataset you have been provided with according to the language you are working with and follow these instructions.
>
> 1. The first goal is to reduce the database to at least 200 claims. For this, first apply a filter on Rating and select only: false, partly false and missing context (or equivalent in the language you are working with).
> 2. Please keep in mind that the fact-checkers providing content to the Google Fact Check Explorer have different ways of naming the ratings but you should find those referring to false, partly false and missing context.
> 3. Then, discard those claims that are Political fact-checks, meaning those claims attributed directly to what a politician said. You can do this by erasing those claims of political fact-checks directly in the database. For example, those of Kamala Harris and Donald Trump:
> 4. If this did not reduce the database to at least 200 claims, please let us know through email.
> 5. Once you have a dataset of between 100 and 200 claims, you can proceed to fill in the fields of Verdict and Relevant Links.
>
>   - **Verdict**: a few sentences long text explaining why the claim is false/partly false, etc. This works as evidence or as the basis

for the rating. It should be as neutral as possible. For example, in this claim: "Pope Francis has been seen partying and drinking alcohol", the verdict would be: "Details in the images (that were circulated online) that may indicate that they have been created with AI tools have been detected. Also, it was proved that the social media profiles disseminating the content were defined as "meme accounts". When discussing images or video, the verdict should be comprehensible even without viewing the media, as in the provided example.

- **Relevant Links**: provide a list of links on which the verdict is based (i.e., containing relevant evidence, they can be information outside of your organization). The relevant evidence should be text-based (i.e., no direct link to video or images). -

6. Please let us know whenever you complete the task or if you have any questions.

B.2 Dataset Analysis

B.2.1 Topic Modeling

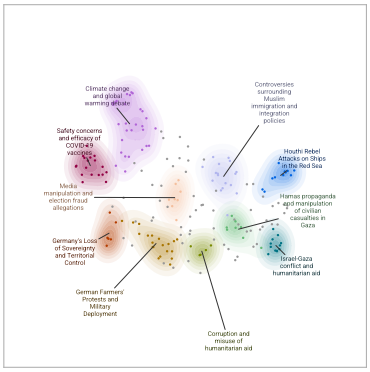
For topic modeling, we implemented the *BERTopic* strategy using `BAAI/bge-m3` as the embedding model. Dimensionality reduction was performed with *UMAP*(`n_neighbors=15`, `n_components=20`, `min_dist=0.0`, `random_state=42`, `metric="cosine"`), while clustering was carried out using *HDBSCAN* (`min_cluster_size=10`, `metric="euclidean"`, `cluster_selection_method="eom"`, `prediction_data=True`). To automatically assign cluster-specific labels, we employed `Llama-3.1-8B-Instruct` for one-shot classification, using the following prompt.[1]

`SYSTEM:` You are a helpful, respectful and honest assistant for labeling topics.
`EXAMPLE PROMPT:` I have a topic that contains the following multilingual documents:
- Traditional diets in most cultures were primarily plant-based with a little meat on top, but with the rise of industrial style meat production and factory farming, meat has become a staple food.
- Meat, but especially beef, is the worst food in terms of emissions.
- Eating meat doesn't make you a bad person, not eating meat doesn't make you a good one.
The topic is described by the following multilingual keywords: 'meat, beef, eat, eating, emissions, steak, food, health, processed, chicken'.
Based on the information about the topic above, please create a short English label for this topic. Make sure you only return the label and nothing more. """
`EXAMPLE PROMPT REPLY:` Environmental impacts of eating meat
`MAIN PROMPT:` I have a topic that contains the following multilingual documents:
[DOCUMENTS]
The topic is described by the following multilingual keywords: '[KEYWORDS]'.
Based on the information about the topic above, please create a short English label

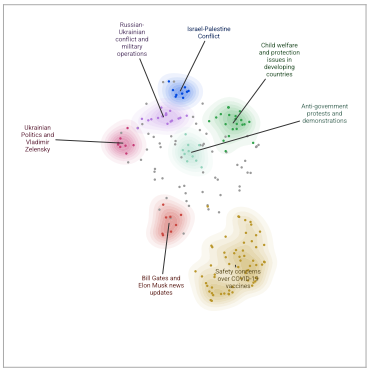
[1] We adapted the prompt provided on the BERTopic documentation https://maartengr.github.io/BERTopic/getting_started/representation/llm.html#llama-2

for this topic. Make sure you only return the label and nothing more.

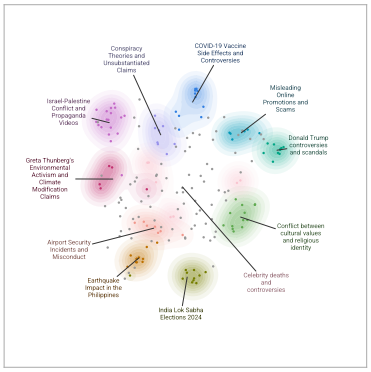
For visualization, we further reduced the embeddings to two components using *UMAP* n_neighbors=15, n_components=2, min_dist=0.0, metric="cosine", random_state=42). In Figure 5.2, we present the topics for all EuroVerdict data. Language-specific topic modelling visualisations are presented in Figure B.1.



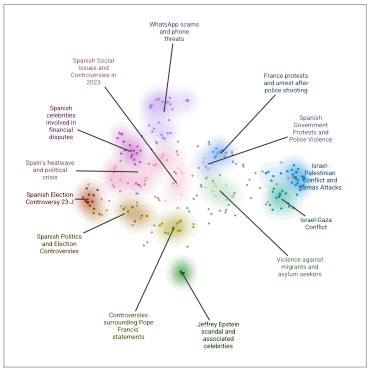
(a) German



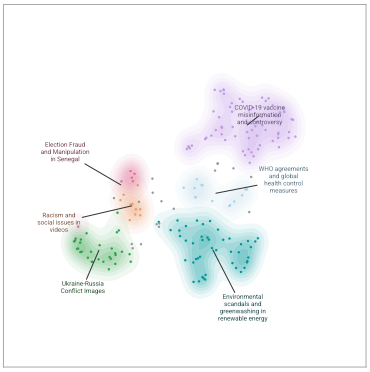
(b) Greek



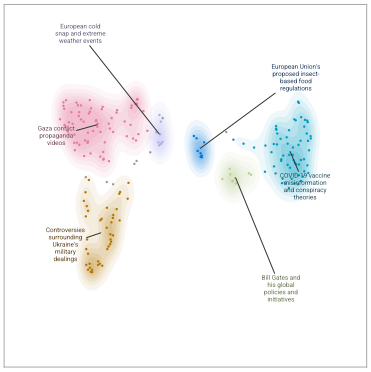
(c) English



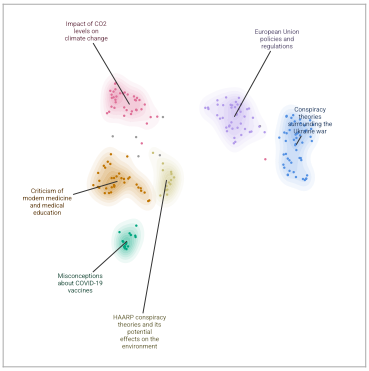
(d) Spanish



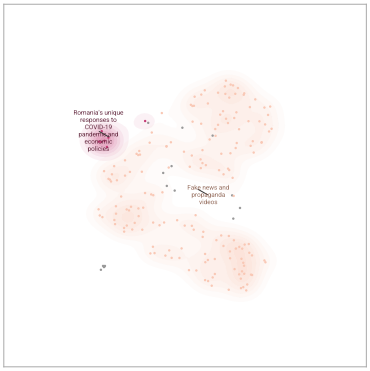
(e) French



(f) Italian



(g) Polish



(h) Romanian

Figure B.1: Topic analysis for each language in EuroVerdict dataset.

B.3 Experimental Design Details

B.3.1 Dataset Partition

For the experiments, we divided EuroVerdict into three subsets: train, evaluation, and test sets. In Table B.1, we report the number of items in each set, both for the entire dataset and for each individual language.

	All	DE	EL	EN	ES	FR	IT	PL	RO
Train	1152	142	138	138	186	135	143	143	135
Eval	245	31	30	30	40	29	31	32	30
Test	245	31	30	30	40	29	31	32	30

Table B.1: Data distribution across train, eval, and test sets.

B.3.2 Expert-Based LLM selection

To assess whether `Llama-31-8B-Instruct` could not only generate grammatically correct text in all eight languages of EuroVerdict but also produce plausible verdicts, we conducted a preliminary zero-shot verdict generation experiment followed by a human evaluation. Similar to the main experiments, we provided the model with a subset of claims in all eight languages along with their corresponding fact-checking articles and instructed it to generate verdicts based on the provided information. The generated verdicts were then evaluated by expert annotators, who were native speakers of the respective languages, and rated on a scale from 1 to 5 for grammatical correctness and soundness (i.e., *how closely the generated verdict resembles a real verdict*). Results showed that across all languages, the generated verdicts were considered both grammatically correct and plausible. Based on these findings, we proceeded to employ the LLM throughout the experiments presented in this work.

B.3.3 Chain of Thought Prompt

Hereafter, we provide the prompt employed in the Chain of Thought configuration.

`SYSTEM:` You will be provided with a misleading claim. Evaluate this claim step-by-step using the given article. To do so, answer the following questions (reasoning). Finally, state whether the claim is true or false (veracity label), and provide a short reply to the claim, providing the reasons why the claim is true or false (justification).
1. What does the claim specifically state?
2. What context or background information in the article is relevant to evaluating the claim?
3. What evidence supports or contradicts the claim?
4. Are there exceptions or scenarios where the claim doesn't apply?
5. To what extent is the claim accurate or misleading? Provide a conclusion with key caveats or nuances.
`USER:` <claim> "`{claim}`" <claim>
<article> "`{article}`" <article>
Reply in `{language}`. Return the response in JSON format: {"claim": "claim", "rea-

soning": [list of the replies to each question], "veracity_label": "TRUE or FALSE", "justification": "why the claim is true or false"} Leave the JSON names in English. Ensure the structure remains consistent throughout.

B.3.4 Fine-Tuning Details

We utilized `Llama-3.1-8B-Instruct` for verdict generation. Four versions of the LMM were fine-tuned, combining English and language-specific prompts with the original fact-checking articles and the English-translated ones. The same hyper-parameters were used across all the fine-tuning, with the only variation being the training data. In particular, we employed the prompt presented in Paragraph 5.4, adding the gold verdict as for the assistant role. The training was performed on an `NVIDIA Ampere A40` GPU with 48GB of memory, applying a 4bit quantization. Low-Rank Adaptation (LoRA) was utilized with a rank of 16, an α value of 16, and a dropout rate of 0. Training parameters included a learning rate of 5×10^{-5}, a training and evaluation batch size of 6, and gradient accumulation steps of 4. The model was trained for 5 epochs, with a weight decay of 0.01 and a warm-up ratio of 0.03. For inference, we employed the checkpoint with the lowest evaluation loss.

B.4 Results Details

B.4.1 Language-Specific Results

In Tables B.3 and B.4, we present the results for all the experiments. For clarity, experiments with articles in the original language are reported in Table B.3, while those with articles automatically translated into English are shown in Table B.4.

B.4.2 Human Evaluation Instructions

Hereafter, we report the task proposed to the human evaluators.

> We prepared a list of Claims and Gold Verdicts.
>
> For each of them, we provide a list of 5 additional verdicts that we ask you to evaluate according to:
>
> - **Alignment**: How much is the current verdict similar to the gold verdict?
> - **Grammaticality**: How much is this verdict written in an understandable way (regardless of being a proper verdict)?
> - **Soundness**: How much does the verdict sound like a well-formed verdict?
>
> Score each of the provided verdicts on a scale from 1 to 5, according to the following guidelines:
>
> **Alignment**
>
> 5. Arguments and conclusions are semantically the same
> 4. Arguments and conclusions are very similar

3. Arguments and conclusions have some discrepancies
2. The conclusion is similar but arguments are completely different (or vice versa)
1. Arguments and conclusions are completely different

Grammaticality

5. Understandable and grammatical
4. Understandable but with few grammatical errors
3. Understandable but with several grammatical errors
2. Difficult to understand with severe grammatical errors
1. Not understandable"

Soundess

5. Very convincing verdict
4. Proper verdict
3. Acceptable Verdict
2. Not a proper verdict
1. Does not seem like a legit verdict"

B.4.3 Human Evaluation Results Details

In Table B.2 we report the scores obtained from the human evaluation of the generated verdicts. Details are provided in Section 5.5.1.

		Overall Mean	Alignment	Understandability	Soundess
Fine-Tuning	OA-EN	4,255	4,118	4,412	4,235
	OA-lang	4,235	4,118	4,471	4,118
	TA-EN	3,159	3,13	3,391	2,957
	TA-lang	3,377	3,435	3,565	3,13
Zero Shot	OA-EN	4,104	4,188	4,312	3,812
	OA-lang	3,863	3,647	4,412	3,529
	TA-EN	3,273	3,409	3,227	3,182
	TA-lang	3,364	3,500	3,500	3,091
CoT	OA-EN	3,725	3,824	3,588	3,765
	TA-EN	2,681	3,000	2,435	2,609

Table B.2: Human evaluation results averaged across all languages. We report the overall score (*Overall Mean*) and scores for *Alignment*, *Understandability*, and *Soundness* across configurations: fine-tuning, zero-shot, CoT, original (*OA*) or translated articles (*TA*), and English (*EN*) or language-specific prompts (*lang*).

		ROUGE-1		ROUGE-L		BertScore		Cosine Similarity	
		EL	**LS**	**EL**	**LS**	**EL**	**LS**	**EL**	**LS**
DE	Zero Shot	0.293	0.288	0.243	0.248	0.752	0.750	0.650	0.654
	Fine-tune	**0.378**	**0.393**	**0.353**	**0.362**	**0.786**	**0.792**	**0.704**	**0.724**
	CoT	0.221	-	0.176	-	0.713	-	0.652	-
EL	Zero Shot	0.294	0.347	0.217	0.257	0.721	0.729	0.599	0.657
	Fine-tune	**0.562**	**0.662**	**0.507**	**0.628**	**0.821**	**0.860**	**0.778**	**0.820**
	CoT	0.267	-	0.180	-	0.694	-	0.626	-
EN	Zero Shot	0.429	0.419	0.305	0.287	0.769	0.758	**0.760**	**0.739**
	Fine-tune	**0.470**	**0.494**	**0.373**	**0.383**	**0.784**	**0.786**	0.723	0.733
	CoT	0.391	-	0.263	-	0.741	-	0.730	-
ES	Zero Shot	0.388	0.340	0.301	0.257	0.758	0.743	0.633	0.580
	Fine-tune	**0.474**	**0.438**	**0.406**	**0.346**	**0.799**	**0.779**	**0.716**	**0.696**
	CoT	0.356	-	0.269	-	0.747	-	0.666	-
FR	Zero Shot	0.307	0.290	0.209	**0.193**	0.724	**0.717**	**0.648**	**0.645**
	Fine-tune	**0.327**	**0.299**	**0.220**	**0.193**	**0.726**	0.714	0.616	0.601
	CoT	0.273	-	0.169	-	0.706	-	0.611	-
IT	Zero Shot	0.288	0.273	0.185	**0.191**	0.722	0.713	**0.694**	0.607
	Fine-tune	**0.293**	**0.298**	**0.201**	0.189	**0.731**	**0.723**	0.671	**0.674**
	CoT	0.250	-	0.166	-	0.708	-	0.611	-
PL	Zero Shot	0.291	0.238	**0.228**	0.176	0.706	0.690	**0.696**	0.609
	Fine-tune	**0.278**	**0.267**	0.210	**0.214**	**0.709**	**0.717**	0.692	**0.725**
	CoT	0.176	-	0.133	-	0.674	-	0.596	-
RO	Zero Shot	0.290	**0.292**	0.198	**0.213**	0.710	0.712	0.640	0.670
	Fine-tune	**0.292**	0.285	**0.201**	0.200	**0.712**	0.712	**0.675**	**0.681**
	CoT	0.202	-	0.139	-	0.685	-	0.657	-
Mean	Zero Shot	0.325	0.311	0.238	0.229	0.734	0.727	0.669	0.642
	Fine-tune	**0.387**	**0.393**	**0.312**	**0.315**	**0.760**	**0.761**	**0.698**	**0.707**
	CoT	0.270	-	0.190	-	0.710	-	0.644	-

Table B.3: Experimental results of `Llama-3.1-8B-Instruct` on the EuroVerdict dataset. We report performance across Zero-Shot, Fine-Tuning (using *articles in the original language*), and Chain of Thought (CoT) settings. Evaluation metrics include ROUGE-1, ROUGE-L, BERTScore, and Cosine Similarity. For each language, results are presented for two prompt types: EL (English-language prompts) and LS (language-specific prompts).

		ROUGE-1		ROUGE-L		BertScore		Cosine Similarity	
		EL	**LS**	**EL**	**LS**	**EL**	**LS**	**EL**	**LS**
DE	Zero Shot	0.264	0.281	0.211	0.240	0.732	0.745	0.641	0.670
	Fine-tune	**0.422**	**0.351**	**0.381**	**0.314**	**0.805**	**0.777**	**0.722**	**0.681**
	CoT	0.200	-	0.159	-	0.701	-	0.634	-
EL	Zero Shot	0.265	0.296	0.190	0.196	0.698	0.701	0.608	0.647
	Fine-tune	**0.343**	**0.349**	**0.280**	**0.284**	**0.731**	**0.730**	**0.695**	**0.744**
	CoT	0.226	-	0.162	-	0.675	-	0.566	-
EN	Zero Shot	0.437	0.422	0.298	0.299	0.766	0.766	0.718	0.742
	Fine-tune	**0.472**	**0.515**	**0.367**	**0.407**	**0.782**	**0.798**	**0.737**	**0.754**
	CoT	0.381	-	0.255	-	0.739	-	0.710	-
ES	Zero Shot	0.345	0.340	0.256	0.257	0.751	0.743	0.675	0.645
	Fine-tune	**0.395**	**0.405**	**0.294**	**0.329**	**0.770**	**0.770**	**0.869**	**0.676**
	CoT	0.316	-	0.232	-	0.734	-	0.603	-
FR	Zero Shot	**0.303**	0.270	**0.213**	0.182	**0.719**	0.710	**0.632**	0.614
	Fine-tune	0.275	**0.307**	0.181	**0.195**	0.712	**0.718**	0.605	**0.637**
	CoT	0.262	-	0.159	-	0.697	-	0.601	-
IT	Zero Shot	0.261	0.289	0.177	0.202	0.710	0.721	0.652	0.652
	Fine-tune	**0.277**	**0.293**	**0.196**	**0.207**	**0.717**	**0.725**	**0.666**	**0.662**
	CoT	0.233	-	0.158	-	0.700	-	0.606	-
PL	Zero Shot	**0.270**	0.209	**0.205**	0.156	**0.705**	0.674	**0.732**	0.637
	Fine-tune	0.232	**0.232**	0.180	**0.178**	0.701	**0.697**	0.703	**0.660**
	CoT	0.207	-	0.171	-	0.672	-	0.589	-
RO	Zero Shot	**0.264**	0.255	0.180	**0.181**	**0.706**	0.699	**0.641**	**0.652**
	Fine-tune	0.260	**0.272**	**0.195**	0.178	0.699	**0.703**	0.636	0.628
	CoT	0.226	-	0.152	-	0.686	-	0.621	-
Mean	Zero Shot	0.302	0.296	0.218	0.216	0.724	0.721	0.664	0.657
	Fine-tune	**0.336**	**0.342**	**0.261**	**0.264**	**0.741**	**0.741**	**0.682**	**0.680**
	CoT	0.258	-	0.183	-	0.702	-	0.616	-

Table B.4: Experimental results of `Llama-3.1-8B-Instruct` on the EuroVerdict dataset. We report performance across Zero-Shot, Fine-Tuning (using *articles translated in English*), and Chain of Thought (CoT) settings. Evaluation metrics include ROUGE-1, ROUGE-L, BERTScore, and Cosine Similarity. For each language, results are presented for two prompt types: EL (English-language prompts) and LS (language-specific prompts).

C

Chapter 6

C.1 Dataset Details

In this work, we employed data from the FullFact and VerMouth datasets (see Chapters 5 and 6). In Table C.1, we detail the distribution of entries across the training, evaluation, and test sets for each dataset.

			Train	Eval	Test
	FullFact		1470	184	174
VerMouth		**SMP**	1470	184	174
	emotions	anger	1265	158	158
		disgust	1339	164	163
		fear	1440	179	171
		happiness	1200	165	149
		sadness	1404	173	171
		surprise	1433	181	170

Table C.1: FullFact and VerMouth data distribution.

C.2 Fact Extraction Module

In order to remove noise from VerMouth's claims, we prepend a fact extraction module before passing the claim to the retriever. To this end, we prompted `Llama-2-13b` and provided an example of the expected output (one-shot). Hereafter, we report the prompt employed:

`SYSTEM:` Extract from the following text the main fact. Remove possible opinions or emotional statements.

Report results in the following format: FACT:[main fact]

Here there is an example:
TEXT: "I just heard about the Covid-19 vaccines & sadly they don't seem to be very effective in preventing the virus. Really disappointing! #vaccineineffective #covid19vaccin "
FACT: "The Covid-19 vaccines offer very little protection against the disease."

`USER:` Now extract the main fact from the following text: TEXT:{claim}

To evaluate the performance of the fact extraction module, we randomly selected 70 instances, evenly distributed across the different claim types. An expert evaluator was provided with a list of claims from the VerMouth dataset along with the corresponding extracted facts generated by the `Llama-2-13b` model. The evaluator was then asked to assess whether the model had successfully identified and extracted the underlying fact. Results show that in only 3 cases (4%), the model failed to extract the fact. Notably, in these instances, the original claims framed the information as an opinion rather than an objective fact. Consequently, the model reproduced the speaker's opinion instead of isolating the factual content. An example is presented below.

> *"As a student, the thought of having more teachers than necessary disgusts me. It's not about quantity, it's about quality education. Let's invest in our teachers and give them the support they need to make a real difference in students' lives. #educationreform #qualityoverquantity"*
>
> *"The speaker believes that investing in teachers and providing them with support is important for quality education."*

C.3 Extra Evidence Extraction Details

In Table C.2, we present detailed information about the additional evidence extracted from FullFact fact-checking articles, used to approximate the realistic scenario in which a gold fact-checking article is not available or does not exist (yet). From the original FullFact fact-checking articles, we removed links to social networks and the source URL of the claim. Indeed, the claims fact-checked by FullFact vary in nature, often originating from social media posts, images, videos, and sometimes misleading headlines. Consequently, the source of the claim might not always provide additional information beyond the claim itself that can be used for verification. Furthermore, even if the claim's source contains extra text, the information can potentially be misleading. Therefore, following our "reliability requirement" we filtered out the claim sources.

C.4 Retrieval Experiments Details

C.4.1 Retrievers Details

For the LLM-Retrieval configuration, we employed `e5-mistral-7b-instruct`, making slight modifications to the original prompt to better align it with our task requirements. The following prompt was employed:

	extra art	extra	words	sent	chunks
all	4093	4	970	38	69412
test	672	4	868	35	9983

Table C.2: Statistics for all additional evidence extracted from FullFact fact-checking articles and the test set used in our experiments. We report the total number of extra evidence documents (*extra art*); the average number of extra documents per fact-checking article (*extra*); the average number of words (*words*) and sentences (*sent*); and the total number of chunks (*chunks*).

Instruct: Retrieve relevant documents to support or refute the given claim.
Query: "{query_str}"

C.4.2 Retrieval Results

In Figure C.1, we report hit_rate, Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP) for the retrieval experiments.

C.5 Generation Experiments Details

C.5.1 Model's instruction

Hereafter, we report the instruction employed for the zero-shot setting. A similar instruction was modified by adding an example from the training sets in the one-shot configuration.

`SYSTEM:` Based on the provided context, respond to the claim, ensuring a thorough explanation. Use only the given context. Reply in no more than three sentences. Avoid mentioning the context in the reply. Match the communication style of the claim and address the possible emotional component present in it, if needed. If the context is insufficient, state that you don't know. Format your response as follows:

Reply: [your_reply]

`USER:` The context information is provided below (in between xml tags).

<context>
{context_str}
</context>

Claim: "{query_str}"

C.5.2 Training Set Creation

In order to fine-tune the LLM for the RAG-based verdict production task, three main elements are needed: a claim, a gold answer, and the context comprising the knowledge needed to reply. In FullFact and VerMouth, the knowledge is present in the form of a fact-checking article. This comes in useful when the entire article is used as a context, but when working with chunks, a proper selection of the most

informative chunks must be performed. To this end, we started from the gold verdicts present in the two aforementioned datasets, and we ranked each article's chunks given the verdict information using a cross-encoder reranking model, i.e., the `BAAI/bge-reranker-large`[1]. Both for the articles and the chunk configurations, we add to the context of each training entry some negative examples, as in testing time

[1] https://hf.co/BAAI/bge-reranker-large

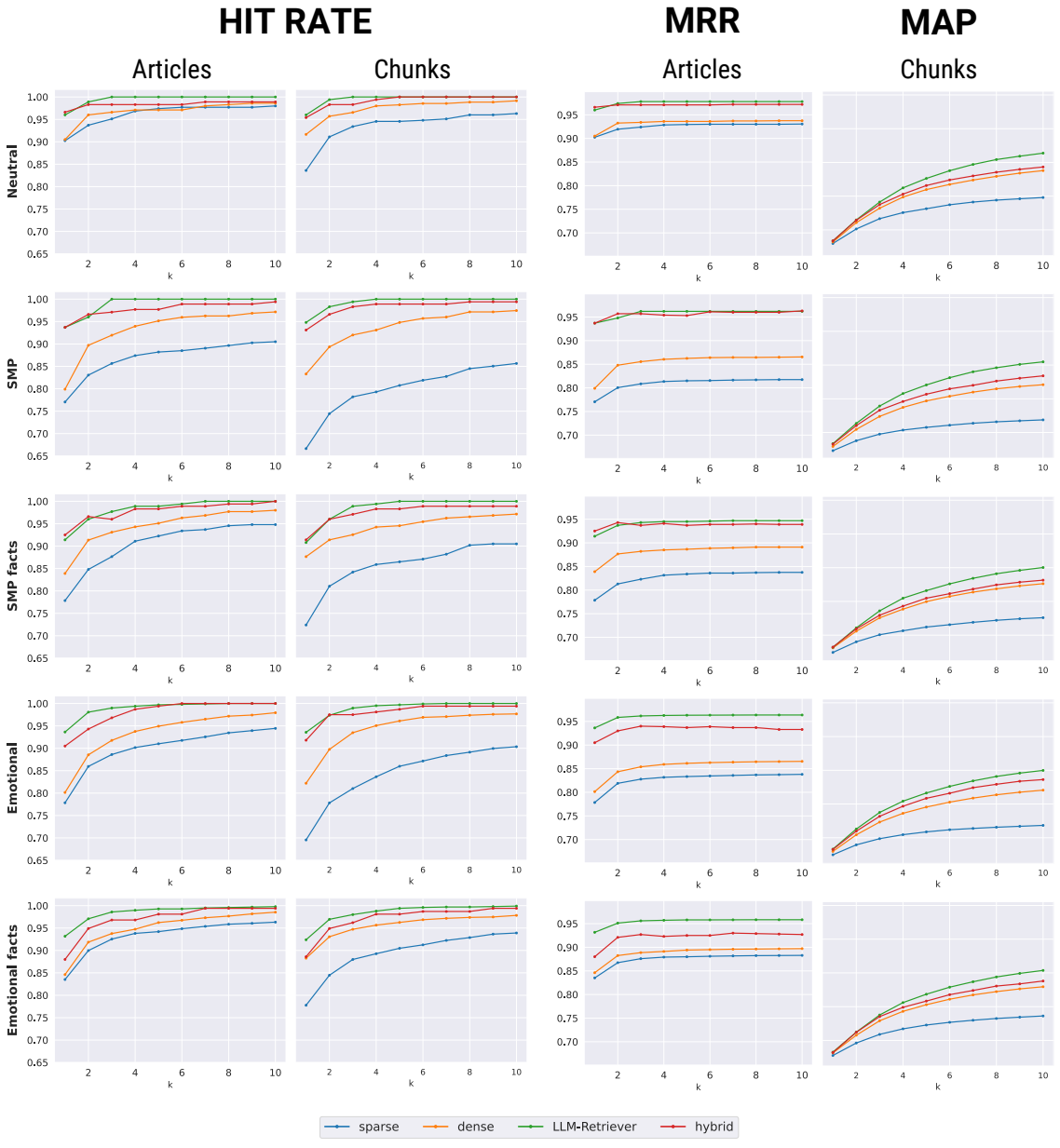


Figure C.1: Retrieval results for each type of retriever (sparse, dense, LLM, hybrid) across Gold_KB$_{art}$ and Gold_KB$_{chunks}$ are presented for all claim styles, both with (SMP Facts, Emotional Facts) and without (neutral, SMP, emotional) claim pre-processing. The metrics reported include hit_rate and MRR for retrieval over Gold_KB$_{art}$, and hit_rate and MAP for Gold_KB$_{chunks}$, for increasing values of retrieved documents/chunks ($k = 1, ..., 10$).

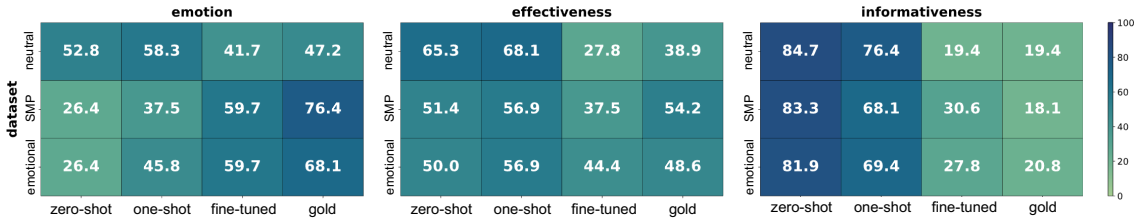


Figure C.2: Complete results for the human evaluation. Each matrix refers to the results obtained for each verdict evaluation aspect. The matrices report how many times, in percentage, the human annotators preferred each of the four generation setups (gold, zero-shot, one-shot, fine-tuning).

the retrieved content might comprise articles or chunks that are not gold. To do so, we employed BM25 for retrieving 10 articles/chunks for each gold verdict and selected the non-gold retrieved context. An example of training input is provided in Table C.3.

C.5.3 Fine-Tuning Details

We fine-tuned the `Llama-2-13b`[2] chat model on different subsamples of training data from the FullFact and VerMouth datasets. From the training dataset, created following the procedure explained in Section C.5.2, we randomly extracted 200 entries each from the FullFact and SMP datasets. For the emotional datasets, we sampled 35 entries for each emotion, totalling 210 training entries. An example of input is shown in Table C.3.

All models were trained on a single Ampere A40 with 48GB memory using the QLoRA strategy (Dettmers et al., 2023), with a low-rank approximation set to 64, a low-rank adaptation set to 16, and a dropout rate of 0.1. Evaluation steps were set at 25, and the batch size was 4. All models were trained for 3 epochs with a learning rate of 10^{-4}.

C.5.4 Generation Results

In Table C.4, we report the complete results for zero-shot and one-shot experiments using chunks and articles as information context.

C.5.5 Generation Examples

In Tables C.5 and C.6, we show examples of generations with claims from both FullFact and VerMouth (anger emotion) datasets. Each table comprises the following information: the claim; the gold verdict; the generations with `Llama-2-13b-chat` model in zero-shot, one-shot, and fine-tuning settings; the relevant evidence retrieved (either chunks, Table C.5, or articles, Table C.6).

C.6 Human Evaluation Details

For the human evaluation of the generated verdict, we enrolled three volunteer evaluators. They were provided with pairs of verdicts (either gold or generated using zero-shot, one-shot, or fine-tuned models) and their corresponding claims. They were

[2] https://hf.co/meta-llama/Llama-2-13b-chat-hf

instructed to assess the best verdict based on three aspects: effectiveness, informativeness, and emotional/empathetic coverage. Hereafter, we list the tasks/questions that evaluators were required to follow when judging the verdict pair.

Informativeness

> Tell which of the two verdicts contains more information supporting its stance.

Emotional Coverage

> Some of the claims can express a variety of emotions. Tell which of the two verdicts better takes into consideration the emotion of the claim by responding with empathy.

Effectiveness

> Which of the two is an overall better verdict (with respect to the claim) that could be used to answer the claim?

Given that the claims presented may cover sensitive subjects, we have incorporated a cautionary note in the task description: *"This task may contain text that some readers find offensive."*. Additionally, we briefed the evaluators on the study's objectives and assured them that all collected data would be anonymized and solely utilized for research purposes.

In Figure C.2 we report the full outcome of the human evaluation, showing the details of the selected verdicts over the three datasets according to emotional coverage, informativeness, and effectiveness.

<s>[INST] «SYS» Based on the provided context, respond to the claim, ensuring a thorough explanation. Use only the given context. Reply in no more than three sentences. Avoid mentioning the context in the reply. Match the communication style of the claim and address the possible emotional component present in it, if needed. If the context is insufficient, state that you don't know. Format your response as follows:

Reply: [your_reply]

«/SYS»

The context information is provided below (in between xml tags).

<**context**>

A meme shared on Facebook features actor John Krasinski in The Office with a whiteboard with edited text, which says: "3 countries refused the covid vaccine", followed by: "Now all 3 of their presidents have died unexpectedly". Beneath the image are the names of the former presidents of Haiti (Jovenel Moïse), Tanzania (John Magufuli) and Zambia (Kenneth Kaunda).

The president did not refuse the Covid-19 vaccines for Zambia. In fact, in March 2021, the Zambian health minister announced plans to vaccinate all over 18s in the country. Similar claims have been fact checked before.

This survey covers households in England and Wales and so does not cover groups (such as those living in student halls of residence), who have "potentially high proportions of drug use", meaning the true figure could be higher. Comparing England & Wales to other countries in Europe is difficult because not all countries have up to date data.

It's correct that cocaine use among 16 to 24 years olds in England and Wales is at its highest level for around a decade. In 2017/18 6% said they had used at least once in the previous year. The claim referred to Britain, but used data covering only England & Wales. We're focusing on England & Wales as data for Scotland and Wales are not available for the most recent year.

There is no evidence to suggest that the death was related to Mr Magufuli's stance on the Covid-19 vaccines. There has been some speculation from Tanzanian opposition leaders, and on social media, that Mr Magufuli's death may have been caused by Covid-19, however this has been discredited. President Kaunda died of pneumonia at a military hospital in Lusaka in June 2021, age 97.

President Magufuli reportedly said that home treatments such as steam inhalation were preferable to "dangerous foreign vaccines", and in February 2021 the country's health minister said that Tanzania had no plans to accept Covid-19 vaccines. Mr Magufuli's successor, president Samia Suluhu Hassan announced that the president's death in March 2021 was due to heart disease.

Arrests have been made but there are still many unknown details about the assasination. There is no evidence to suggest that there is a link to the lack of progress made regarding Haiti's vaccine roll out.

President Moïse did not explicitly refuse all of the Covid-19 vaccines, but the country did initially refuse the AstraZeneca vaccine due to safety concerns. There have also been administrative and supply difficulties around the arrival of vaccinations to Haiti, and concerns about the logistics of organising the vaccine roll out. Mr Moïse was fatally shot at home on 7 July 2021.

The posts suggest that the death of these three men is linked to a refusal of the Covid-19 vaccines. There is no evidence to support this claim. Haiti is eligible for provision of Covid-19 vaccines through the Covax programme, but roll out has been slow, with Haiti yet to give any doses to residents, and only receiving its first vaccines this week.

The Royal College of Obstetricians & Gynaecologists (RCOG) says that there is "no evidence" to suggest that the Covid-19 vaccines affect fertility. Similar statements have been issued by the British Fertility Association, the NHS and Public Health England. The RCOG also states that there is no biologically plausible mechanism by which the vaccine would cause fertility problems.

</**context**>

Claim: "The presidents of Haiti, Tanzania, & Zambia all passed away suddenly & unexpectedly. Reports suggest it may be connected to their refusal to accept the Covid-19 vaccines for their countries. This is a tragic loss for the world & for the people of those countries. #vaccinesaveslives #RIP" [/INST]

Reply: "I'm sorry to hear about the recent passing of the presidents of Haiti, Tanzania, and Zambia. However, it's important to clarify that the president of Zambia did not refuse vaccines and passed away from pneumonia at a ripe old age, while the president of Tanzania passed away from a heart condition. The assassination of the president of Haiti is a tragic event, but there is no evidence linking it to vaccine refusal. Let's honor their legacies and focus on ways to prevent further loss of life."

Table C.3: Training example for `Llama-2-13b` model. The positive passages are highlighted in green, while negative in red.

		Articles				Chunks			
Zero-shot									
model	**dataset**	**ROUGE-LSum**	**BARTScore**	**SummaC**	**GoldSim**	**ROUGE-LSum**	**BARTScore**	**SummaC**	**GoldSim**
mistral-7b-v0.1	fullfact	0,216	-1,948	0,381	-3,101	0,202	-1,992	0,361	-3,112
	SMP	0,185	-2,153	0,333	-2,990	0,190	-2,207	0,351	-3,060
	emotional	0,174	-2,389	0,322	-2,953	0,188	-2,411	0,327	-2,990
mistral-7b-v0.2	fullfact	0,130	-2,315	0,331	-3,084	0,144	-2,251	0,334	-3,101
	SMP	0,140	-2,516	0,327	-3,144	0,145	-2,453	0,335	-3,078
	emotional	0,131	-2,803	0,312	-3,134	0,145	-2,789	0,317	-3,143
llama3-8b	fullfact	0,050	-3,143	0,310	-3,516	0,035	-3,439	0,319	-3,648
	SMP	0,047	-3,330	0,295	-3,644	0,032	-3,461	0,328	-3,715
	emotional	0,054	-3,142	0,282	-3,466	0,041	-3,217	0,305	-3,575
llama2-7b	fullfact	0,168	-1,979	0,330	-2,914	0,181	-1,988	0,330	-2,932
	SMP	0,160	-2,118	0,321	-2,861	0,195	-2,022	0,330	-2,898
	emotional	0,159	-2,253	0,311	-2,755	0,197	-2,171	0,318	-2,781
llama2-13b	fullfact	0,176	-1,714	0,353	-2,787	0,195	-1,718	0,355	-2,811
	SMP	0,169	-1,849	0,331	-2,714	0,186	-1,879	0,352	-2,752
	emotional	0,156	-2,118	0,323	-2,691	0,183	-2,058	0,338	-2,754
One-shot									
mistral-7b-v0.1	fullfact	0,173	-2,186	0,336	-3,073	0,180	-2,224	0,341	-3,166
	SMP	0,160	-2,369	0,339	-3,043	0,161	-2,523	0,325	-3,018
	emotional	0,145	-2,586	0,307	-2,947	0,162	-2,664	0,302	-3,012
mistral-7b-v0.2	fullfact	0,134	-2,347	0,321	-3,177	0,136	-2,393	0,320	-3,206
	SMP	0,123	-2,630	0,301	-3,160	0,130	-2,623	0,304	-3,117
	emotional	0,116	-2,865	0,318	-3,138	0,125	-2,871	0,321	-3,148
llama3-8b	fullfact	0,133	-1,960	0,351	-3,003	0,135	-1,905	0,344	-2,988
	SMP	0,138	-2,322	0,323	-3,011	0,137	-2,377	0,321	-3,013
	emotional	0,131	-2,469	0,328	-2,927	0,134	-2,485	0,334	-2,902
llama2-7b	fullfact	0,171	-2,091	0,327	-2,977	0,197	-2,016	0,325	-2,889
	SMP	0,158	-2,408	0,305	-2,922	0,181	-2,467	0,304	-2,911
	emotional	0,147	-2,561	0,317	-2,875	0,191	-2,549	0,306	-2,869
llama2-13b	fullfact	0,185	-2,057	0,322	-2,942	0,199	-1,992	0,309	-2,840
	SMP	0,154	-2,355	0,314	-2,920	0,172	-2,235	0,316	-2,803
	emotional	0,139	-2,523	0,314	-2,839	0,152	-2,441	0,303	-2,737

Table C.4: Complete results for each model tested on the three datasets in zero-shot and one-shot settings. Results for both chunks and article configurations are reported.

claim

Research says that nearly half of people who catch Covid develop depression afterwards.

gold verdict

This is not what the research says. It measured rates of depression among Covid survivors—but not whether they developed depression after catching the disease.

zero-shot

The claim is partially accurate but misleading. The study in question found that 48% of participants experienced moderate or severe depression, but it did not establish a causal link between Covid-19 and depression. The study also did not compare depression rates among those who had Covid with those who had not, nor did it consider the normal background rate of depression in Bangladesh. Additionally, the study did not find that Covid-19 directly causes depression. Instead, it suggests that people suffering from depression may be more likely to experience long-term symptoms after recovering from Covid-19.

one-shot

The claim is misleading and not entirely accurate. The research in question found that nearly half of the participants reported experiencing depression "alongside" Covid, but it does not necessarily imply that Covid caused the depression. The study did not compare the prevalence of depression among those who caught Covid to those who did not, nor did it account for the normal background rate of depression in Bangladesh. Additionally, the study did not find any causal link between Covid and depression. The researchers themselves noted that the findings may not reflect the impact of Covid-19, especially for depression, which may have developed prior to the pandemic. It is important to be cautious when interpreting the results of scientific studies and not draw conclusions that go beyond the evidence.

fine-tuning

The study found that 48% of people who had Covid had symptoms of moderate or severe depression. But the researchers themselves say that this does not necessarily reflect the impact of Covid, and that it may have developed before the pandemic.

1 - The study in question explicitly said that its findings didn't necessarily reflect the impact of Covid-19. The newspaper said: "Nearly half of people struck down with Covid-19 also develop depression, according to new research."The research in question surveyed the mental health of 1,002 people in Bangladesh who had tested positive for Covid at some point in the past.
2 - These people were a self-selected sample, meaning they heard about the survey and wanted to take part, so they might not be representative of people as a whole. The study found that about 48% had moderate or severe depression. However, it does not tell us that they "go on" to develop depression as a result of catching Covid, or even that they developed it after catching Covid.
3 - One of the study authors says it describes people suffering from depression "alongside" Covid. The researchers themselves say in the paper: "The findings may not reflect the impact of COVID-19, especially for depression, which may have developed prior to the COVID-19 pandemic."There is evidence that rates of depression have increased in Great Britain and elsewhere during the pandemic.
4 - The Telegraph has muddled up the meaning of a scientific study on depression among people who have had Covid-19.
5 - The research doesn't compare its findings with depression rates among these people before they got the disease, nor with people who hadn't been infected at all, nor with the normal background rate of depression in Bangladesh (which may have risen during the pandemic).A press release from Anglia Ruskin University, which worked on the research, does not say that Covid was found to cause depression.
6 - There is other evidence that people may be at higher risk of developing mental health problems after recovering from Covid. The extent of this and why it might happen are not yet fully understood, however.
7 - 26% had symptoms of moderate or severe levels of depression with no real change across the first six weeks, which may be what the "one in four" claim was meant to refer to. Feelings of defeat and entrapment decreased over time, while positive wellbeing increased. Although the findings of the study were not as dramatic as claimed in the video, researchers said there was cause for concern.
8 - So Mr Hancock was right to say that one in 20 people would have long-term symptoms after a Covid infection—if you define "long-term" as more than eight weeks. However, he was wrong to compare this to the one-in-10 figure for people under 50, because this describes those who suffered from symptoms that lasted more than four weeks—rather than eight. As the research paper explains,
9 - Mr Hancock was also wrong to say that there was "no correlation" between the severity of the initial illness and the length of the symptoms afterwards. The research paper says people with long Covid were "more likely to have required hospital assessment in the acute period." Overall, 13.9% of the confirmed cases in the study had visited hospital. However, among those reporting symptoms for at least four weeks this rose to 31.5%,
10 - and among those reporting symptoms for at least eight weeks it was 43.9%. The app data also suggested that "Individuals reporting more than 5 symptoms in the first week (the median number reported) were significantly more likely to go on to experience LC28 symptoms lasting at least four weeks."In short, more severe cases of Covid seemed more likely to result in long Covid afterwards.

Table C.5: Example of generation using FullFact claim, e5-mistral as a retriever and Llama-2-13b-chat for the generation of the verdict. In green are highlighted the gold chunks retrieved.

claim

I'm fuming over the fact that the worst violent and sexual criminals are let off the hook halfway through their sentence. How is this justice? I feel outraged and betrayed by the justice system." #JusticeForVictims

gold verdict

I understand your frustration and anger towards the justice system. It is important to note that judges in England and Wales can already choose to keep dangerous offenders in prison for up to two thirds of their sentence. I hope this information provides some clarity. #JusticeForVictims

zero-shot	one-shot	fine-tuning
The claim is partially accurate but misleading. The study in question found that 48% of participants experienced moderate or severe depression, but it did not establish a causal link between Covid-19 and depression. The study also did not compare depression rates among those who had Covid with those who had not, nor did it consider the normal background rate of depression in Bangladesh. Additionally, the study did not find that Covid-19 directly causes depression. Instead, it suggests that people suffering from depression may be more likely to experience long-term symptoms after recovering from Covid-19.	The claim is misleading and not entirely accurate. The research in question found that nearly half of the participants reported experiencing depression "alongside" Covid, but it does not necessarily imply that Covid caused the depression. The study did not compare the prevalence of depression among those who caught Covid to those who did not, nor did it account for the normal background rate of depression in Bangladesh. Additionally, the study did not find any causal link between Covid and depression. The researchers themselves noted that the findings may not reflect the impact of Covid-19, especially for depression, which may have developed prior to the pandemic. It is important to be cautious when interpreting the results of scientific studies and not draw conclusions that go beyond the evidence.	The study found that 48% of people who had Covid had symptoms of moderate or severe depression. But the researchers themselves say that this does not necessarily reflect the impact of Covid, and that it may have developed before the pandemic.

At the Conservative party conference last week, justice secretary Robert Buckland pledged to make it mandatory for the most serious violent and sexual offenders to serve two thirds of their sentence in prison (and not half the sentence as is normal). The Ministry of Justice told us this would apply to "the most serious sexual and violent offenders where the offence carries a maximum life sentence." However, to suggest that the most serious criminals are automatically released from prison halfway through their sentences obscures the fact that judges already have discretion to keep serious criminals in prison for longer. There are various sentences a judge can hand down and in most cases sentences are non-custodial (where no time is served in prison). By far the most common sentence for crimes in England and Wales is a fine, but what's of interest here are sentences which carry mandatory prison time. Typically in these cases a criminal will be given a standard determinate sentence. This usually requires them to spend half of their sentence in prison and the other half on license in the community, supervised by the probation service. For example, a standard two year sentence would involve one year in prison and one year on license. Being on licence means you can be recalled to prison if you breach the terms of your licence. As well as standard sentences, judges in England and Wales can hand down what are called 'extended determinate sentences' to criminals who commit any of over 100 serious offences. The judge can make this decision if: These offenders are either entitled to be released two thirds of the way through their sentence, or can apply for parole at that point. Life sentences work slightly differently. With a life sentence a criminal is required to spend a minimum time in prison and is then able to apply for parole. If they are released, they remain on license for the rest of their life. Compared to existing extended sentences, the Conservatives' proposal appears to apply to criminals who commit a slightly different group of offences (those that carry a maximum of life rather than this list of serious offences). There is also apparently no requirement for a judge to determine if a criminal poses a risk to the public when giving this new kind of sentence. While the current sentencing procedure does not dramatically change the ability to put serious criminals in prison for two thirds of their term, it would, in practice, significantly increase the number of criminals receiving two thirds sentences. [...]

Table C.6: Example of generation using VerMouth anger claim, e5-mistral as a retriever and Llama-2-13b-chat for the generation of the verdict. The article has been cut for space reasons. The complete article's text can be found at https://fullfact.org/crime/extended-sentences/

D

Chapter 7

D.1 Dataset Creation Details

D.1.1 Category Assessment

In Table D.1 we report how the heterogeneous categories scraped directly from the misleading websites were divided into the four macro-categories of *scienza* (science), *salute* (well-being), *ambiente* (environment), *economia* (economy).

D.1.2 Annotation Guidelines

Three components of our datasets were subject to human intervention to: (i) determine if the headline was clickbait, (ii) identify the related article's spoiler, that is, the information required to satisfy the curiosity gap within the headline, and (iii) revise the headline to include the spoiler information, thereby neutralizing it. During all three annotation stages, we employed a machine-human collaboration to expedite the work of annotators. The annotators received both a score indicating how much the headline was clickbait and automatic ChatGPT `gpt-3.5-turbo-0125` generated suggestions for the spoilers and the neutralized versions of the headlines. Below, we have outlined the annotation guidelines that the annotators were to follow.

Clickbait labelling In order to select the clickbait headlines present in the scraped data, the annotators were provided with specific guidelines. Table D.2 provides the main key points taken into consideration in order to label the data.

Spoiler post-editing For the post-editing of the spoiler, the annotator was required to spot the information gap in the headline and to check if the generated spoiler was providing that information by checking the related article. If the model failed to find the proper spoiler, the annotator had to rewrite it, sticking as much as possible to the document's text. If the spoiler was correct but added extra info, the annotator had to keep that extra information only if those were essential for having a complete headline. If the spoiler was correct, then the annotator could leave it as

it was.

Neutralised Clickbait Post-Editing The annotator was required to check if the neutralised forms comprise both the headline and the spoiler information. If the spoiler was very long (e.g., a long listing), then the annotator had to summarise the spoiler as much as possible, aiming to embed in the final novel headline enough information to reduce or remove the information gap. If the model failed at addressing the spoiler information in the neutralised version of the headline, then the annotator had to manually add it. Moreover, the annotator was required to remove sensationalist tones as much as possible if this tone was still creating useless curiosity in the reader.

D.1.3 Author Component Instruction

Hereafter, we provide the instruction employed to automatically generate spoilers and the neutralised versions of the clickbait headlines through ChatGPT `gpt-3.5-turbo-0125`.

I have a clickbait headline and its corresponding article, both written in Italian. The clickbait headline typically omits key information to create a curiosity gap for the reader. Your task is to extract this missing information, known as a "spoiler," from the article's text. The spoiler can be a single keyword, a short text passage, or a list of keywords. Once you have identified the spoiler, rewrite the clickbait headline by incorporating this information to eliminate the curiosity gap. The output must be in JSON format and written in Italian. The JSON should include two entries: one called "spoiler" that contains the extracted spoiler(s), and another called "new_headline" that has the revised headline.

Example Input:

Clickbait headline: "Questo attore ha fatto qualcosa di incredibile sul set di un famoso film!" Article: "Durante le riprese del film 'Il Gladiatore', l'attore Russell Crowe ha deciso di fare un gesto di grande generosità donando una parte significativa del suo stipendio al fondo per i membri della troupe."

Example Output:

{"spoiler": "Russell Crowe ha donato una parte significativa del suo stipendio al fondo per i membri della troupe", "new_headline": "Russell Crowe ha fatto qualcosa di incredibile sul set di 'Il Gladiatore': ha donato una parte significativa del suo stipendio al fondo per i membri della troupe"}

Please ensure the output is formatted in JSON as specified and that all content is in Italian.

Now do it for the following headline.

Clickbait headline: "{headline}"

Article:"{article}"

D.2 Additional Dataset Details

D.2.1 Dataset Visualisation

Figure D.1 shows a frequency-based visualization of the dataset. It considers the frequency of appearance of relevant uni and bi-grams for both the clickbait and non-clickbait categories. The figure shows common strategies that are frequent in clickbait content, such as the use of "ecco cosa" (*this is what*) or "quali sono" (*what are*)

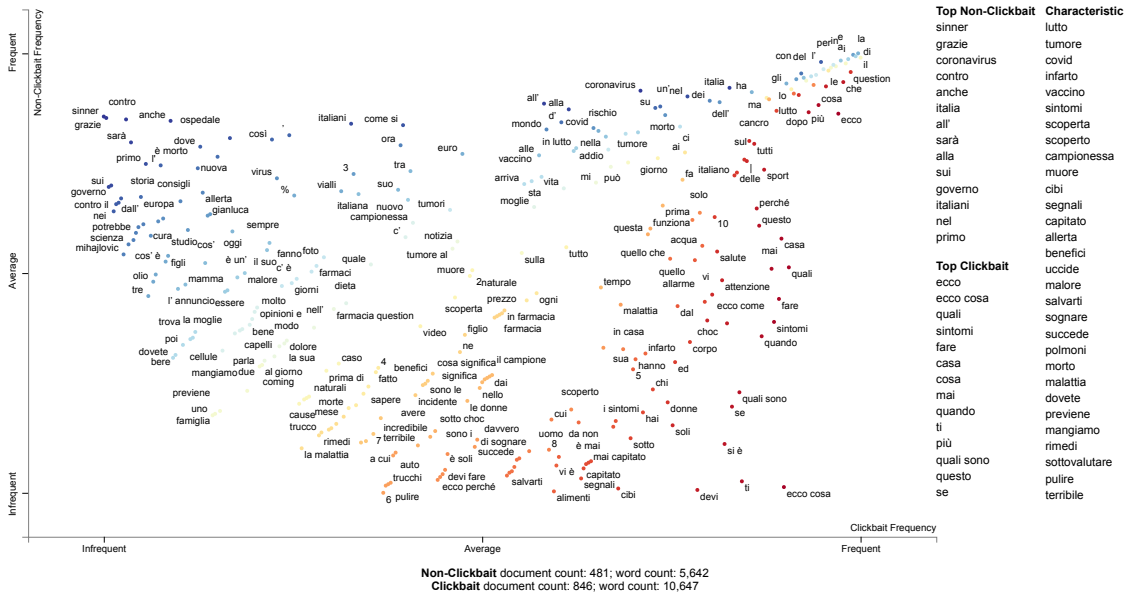


Figure D.1: Frequency of words for both clickbait and non-clickbait categories. On the right, the most frequent words for each class, and both (Characteristic).

that can be seen in the lower right part. An interactive version of the graph can be accessed at the following link https://github.com/oaraque/ClickBaIT/blob/main/docs/clickbait.html.

D.2.2 Dataset Excerpt Translation

Table D.3 includes the English translations for the Italian examples presented in Table 7.1.

D.3 Experimental Design Details

D.3.1 Question Generation

Questions were generated with ChatGPT `gpt-3.5-turbo-0125` using the following prompt:

> You will be provided with a clickbait headline written in Italian. Your task is to generate a question that addresses any missing or vague information in the headline. Here are some examples:
> Headline: Si chiama la benedizione di Dio: rimuove l'alta pressione, il diabete e il grasso nel sangue
> Question: Che cosa viene chiamato 'benedizione di Dio'?
> Headline: "Emorragia cerebrale". Italia in apprensione per il suo campione: ricoverato in condizioni gravissime
> Question: Chi è il campione?
> Please generate the question in Italian, ensuring it seeks to clarify the ambiguous or incomplete details present in the headline.

D.3.2 Spoiler Generation

For the zero-shot spoiler generation task, we employed the following prompt:

> Ti verranno forniti un titolo clickbait e il suo articolo corrispondente. Il titolo clickbait di solito omette, o non esplicita, informazioni chiave per creare curiosità nel lettore. Estrai dall'articolo le informazioni mancanti o vaghe nel titolo che servono per colmare questa curiosità. La risposta può essere un messaggio estremamente conciso oppure un elenco. Formatta la risposta nel seguente modo. "Risposta: <output>"
> Titolo: {headline}
> Articolo: {article}

The same instruction was employed with the fine-tuned model. For the few-shot generation of the spoiler, we enriched the instruction with two examples. When casting spoiler generation as a Question Answering task, the following instruction was employed:

> Ti verrà fornita una domanda e un documento. Trova nel documento le informazioni per rispondere alla domanda. La risposta può essere un messaggio conciso oppure un elenco. Formatta la risposta nel seguente modo. "Risposta: <output>"

D.3.3 Fine-Tuning Details

The `LLaMAntino-3-8B` (Polignano et al., 2024) model underwent training on a single Ampere A40 GPU with 48GB of memory, employing the QLoRA strategy with a low-rank approximation of 64, a low-rank adaptation of 16, and a dropout rate of 0.1. It was set to evaluate every 50 steps, with a batch size of 4, across 3 epochs, using a learning rate of 10^{-4}.

In the clickbait detection experiments, the `DistilBERT` and `Llama3-8b` models have been fine-tuned on the same GPU. The DistilBERT model has been trained on 10 epochs with a learning rate of $2 \cdot 10^{-4}$. For the Llama3 model, we have used QLoRa with the same characteristics as described above, trained on two epochs, with a learning rate of $2 \cdot 10^{-4}$.

D.3.4 Neutralised Clickbait Generation

The following system prompt (enriched with three examples) has been utilised with LLaMAntino-3-8B:

> Ti verranno forniti due testi: un titolo clickbait e un testo, chiamato spoiler, che contiene le informazioni mancanti nel titolo. Il tuo compito è di riscrivere il titolo clickbait integrando le informazioni dello spoiler. Il nuovo titolo deve essere informativo, privo di toni sensazionalistici e breve. Se lo spoiler contiene tante informazioni, puoi riassumerle in concetti più generali.
>
> Titolo: {headline}
> Spoiler: {spoiler}

scienza	insetti, animali, AI, scienza, smartphone, Spazio, tecnologia, TECNOLOGIE, SCIENZA, ufo, biochimica, eclissi, bomba atomica, terra piatta, idroelettrico, temperatura, coltivazione, robot, fisica quantistica, macchie solari, ricerca, vulcano, titanio, universo, fotovoltaico, intelligenza, iPhone, hacker, microonde, motori di ricerca, onde elettromagnetiche, tecnologia, sole, scienza, radioterapia, pesticidi, armi chimiche, comete, case farmaceutiche, psichiatria, smartphone, formiche, elettrodomestici, solare, macrobiologi, mondo, lampadine a basso consumo, tecnologia, scienze-e-tech, scienza, scienza, innovazione, scienza, tecnologia-2, animali intelligenti, funzione cognitiva, microchip, cani, samsung, wi fi, tecnologia-e-tv, SCIENZE, TECNOLOGIA, bioetica, biologia, fisica, covid, coronavirus
salute	Salute, CORONAVIRUS, VAIOLO SCIMMIE, TUBERCOLOSI, SALUTE, SCABBIA, AIDS, salute, hiv, cocaina, antidepressivi, veleni, infezioni, carne, tabacco, infibulazione, fluoro, alcool, alimentari, aids, antibatterico, dieta, insetticida, cibo, benessere, farmaci, digitopressione, caffè, sigarette, ministero della salute, autismo, limoni, cure naturali, paracetamolo, cancro, antiossidante, droga, olio, medicina alternativa, fragole, vegetariano, eroina, dislessia, veleno, zenzero, virus, psicologia, biologico, magnesio, frutta, psicofarmaci, pollo al cloro, fiori di bach, medico, sonno, birra, vitamina e, ulivi, proteine, stress, banana, pensieri negativi, tumori, benzodiazepine, latte, miele, cuore, epilessia, longevità, marijuana, diabete, sale, ibernazione, vecchiaia, fegato, vegan, prevenzione, dentifricio, cervello, sistema immunitario, sodio, suicidio, rimedi naturali, maltempo, canapa, pillola, mal di gola, depressione, psiche, alimentazione, ebola, aspartame, dentifricio senza fluoro, tiroide, mangiare, cure proibite, Alzheimer, smog, gas, malattie, calamità, mammografia, verdura, aloe, masticazione, farmaco, igiene, batteri, medicina, vitamina c, epatite c, forfora, energia, vaccini, ormoni, flora batterica, sorbitolo, antibiotici, piedi, obesità, arsenico, cortisolo, chemioterapia, contraccezione, Neurotrasmettitori, semi, melograno, celiachia, Coca cola, salute-benessere, salute, salute-e-benessere, bellezza, dimagrante, benessere, salute-benessere, rimedi-naturali, pianeta-mamma, grano antico, acqua ossigenata, alimetnazione, ansia, dentisti, curcuma, casa-e-cucina, hobby-e-sport, SPORT, crescita-consapevolezza, la-salute-che-viene, sport, stile-di-vita, consigli, lifestyle, pomodori
ambiente	Cambiamenti climatici, energia, energia elettrica, Natura, AMBIENTE, ECOLOGIA, global warming, geoingegneria, alberi, pianeta terra, natura, inquinamento, mare, terra, manipolazione climatica, clima, rinnovabili, Dissesto idrogeologico, ecologia, ambiente, green, ambiente-attuale, ecologia, salute-benessere, natura, ambiente, METEO, tempesta solare, astronomia, acido
economia	affari-online, economia, ECONOMIA, consumi-risparmi, microchip r-fid, bollo auto, tasso d'interesse, finanza, bollette, banche, profitto, spese, economia-finanza, economia, economia, economia-dellanima, fisco-e-tasse, economia, economia, economia, economia-e-finanza

Table D.1: Split of the categories into the four macro-categories.

Characteristic	Original example (IT)	Translated example (EN)
Lack of essential information, i.e., the subject the article is talking about	*"Ora riposa in pace". Calcio in lutto, morto uno dei grandi protagonisti dell'Italia*	*"Now rest in peace". Football in mourning, one of Italy's great protagonists dead*
Sensationalist tone	*Fan ubriaca le salta addosso sul palco. La sua reazione è incredibile e sconvolge tutti i presenti*	*Drunk fan jumps on her on stage. Her reaction is incredible and shocks everyone present*
Questions raised but answered in the article body	*Tratti della nostra colonna: quali sono? Come evitare lesioni?*	*Traits of our column: what are they? How to avoid injuries?*
Enumeration of elements	*10 cibi per sbarazzarsi del gonfiore di stomaco e pancia*	*10 foods to get rid of bloated stomach and tummy*
Use of capitalization	*INFARTO: sopravvivere quando si è soli. Hai solo 10 secondi per salvarti. Ecco cosa devi fare:*	*HEART ATTACK: surviving when alone. You only have 10 seconds to save yourself. Here's what you have to do:*
Introduction of the content without actually giving the information	*Zanzare, ecco come eliminarle senza insetticidi*	*Mosquitoes, this is how to eliminate them without insecticide*
Use of quotations that do not give information	*Omicron, Ilaria Capua: "Ecco perché i vaccinati si infettano di più rispetto a prima"*	*Omicron, Ilaria Capua: "This is why the vaccinated get more infected than before"*

Table D.2: Key points used for the annotation of the dataset. Please note that some instances can exemplify more than one point.

Category	Headline	Article	Clickbait	Spoiler	Neutralised title
Health	Fruit or flower? Tasty and attractive, a celebrity on our tables, we reveal who she is	We all know it, inevitable on our tables, world-famous, but mysterious is its nature, fruit to enjoy or flower to decorate?	True	The strawberry	Strawberry: tasty and attractive, a celebrity on our tables
Science	Self-repairing metal discovered. Scientists astounded	The recent experiment revealed an extraordinary phenomenon...	True	Platinum	The metal that repairs itself: platinum
Health	A disease that affects 500,000 people	We are talking about a chronic immune-mediated systemic disease that affects about 1.8 million patients...	True	Psoriasis affects about 500,000 people	Psoriasis: a disease that affects about 500,000 people in Italy
Environment	Mosquitoes, here's how to get rid of them without insecticides	With the arrival of hot weather, mosquitoes also make their way into our homes or gardens...	True	To eliminate mosquitoes from your home once and for all, you should buy a bat	Mosquitoes, here's how to get rid of them without insecticides: just buy a bat

Table D.3: Translated from the original Italian. An excerpt of the presented dataset showing the most relevant fields. Article bodies are shortened for space reasons.