UNIVERSIDADE FEDERAL FLUMINENSE

VÍTOR NASCIMENTO LOURENÇO

Towards Trustworthy and Explainable Automated Fact-Checking: from Multimodal Classification to Knowledge-Grounded Reasoning

NITERÓI

2026

VÍTOR NASCIMENTO LOURENÇO

Towards Trustworthy and Explainable Automated Fact-Checking: from Multimodal Classification to Knowledge-Grounded Reasoning

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação.

Orientador:
PROF. ALINE MARINS PAES CARVALHO

Coorientador:
DR. TILLMAN WEYDE

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

L892t Lourenço, Vítor Nascimento
Towards Trustworthy and Explainable Automated Fact-Checking: from Multimodal Classification to Knowledge-Grounded Reasoning / Vítor Nascimento Lourenço. - 2026.
181 f.

Orientador: Aline Marins Paes Carvalho.
Coorientador: Tillman Weyde.
Tese (doutorado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2026.

1. Checagem de fatos. 2. Aprendizado de máquina. 3. Inteligência artificial. 4. Produção intelectual. I. Carvalho, Aline Marins Paes, orientadora. II. Weyde, Tillman, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

VÍTOR NASCIMENTO LOURENÇO

Towards Trustworthy and Explainable Automated Fact-Checking: from Multimodal Classification to Knowledge-Grounded Reasoning

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação.

Aprovada em JULHO de 2026.

BANCA EXAMINADORA

Prof. ALINE MARINS PAES CARVALHO - Orientador, UNIVERSIDADE FEDERAL FLUMINENSE

Dr. TILLMAN WEYDE - Co-orientador, CITY ST GEORGE'S, UNIVERSITY OF LONDON

Prof. CLÁUDIO CAMPELO, UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

Prof. FLAVIA BERNARDINI, UNIVERSIDADE FEDERAL FLUMINENSE

Prof. DANIELA BARREIRO CLARO, UNIVERSIDADE FEDERAL DA BAHIA

Prof. FABRÍCIO BENEVENUTO, UNIVERSIDADE FEDERAL DE MINAS GERAIS

Prof. LEANDRO SANTIAGO DE ARAÚJO, UNIVERSIDADE FEDERAL FLUMINENSE

Niterói
2026

"To boldly go where no one has gone before."

Capt. Jean-Luc Picard

Acknowledgements

To Professor Aline Paes, my advisor, who has believed in my potential since my undergraduate studies, providing unwavering support and invaluable guidance throughout my career. Her mentorship shaped not only this thesis but my scientific vision and the principles I carry as a researcher; if I have seen further, it is by standing on her shoulders.

To Dr. Tillman Weyde, my co-advisor, for his careful guidance, technical insight, and steady support throughout this research. He placed his trust in me and opened the doors for me to work abroad, welcoming me into the city I have since come to call home.

To Professors Cláudio Campelo, Flavia Bernardini, Daniela Barreiro Claro, Fabrício Benevenuto, and Leandro Santiago de Araújo, for agreeing to serve on the Examining Committee and for their insightful contributions.

To all the friends, colleagues, and scientists from whom I had the pleasure to learn so much over these years, and who helped me grow as a professional and as a researcher.

To my parents, Izabel and Cleber, and my brother, Bernardo, for their many sacrifices and for all the personal support that allowed me to persevere towards my goal, even from afar.

To my partner, Júlia, who chose to support me in my career and has stood by me as we build our home and family overseas, for her patience and for the many sacrifices that made this journey possible.

To each of you, I am deeply grateful.

Thank you.

Resumo

A disseminação de desinformação nas redes sociais tem superado a capacidade de checagem de fatos conduzida manualmente por humanos, motivando o desenvolvimento de sistemas de verificação automatizada de fatos (AFC, do inglês *Automated Fact-Checking*) capazes de escalar com o volume e a velocidade do conteúdo *online*. No entanto, apesar dos avanços recentes em processamento de linguagem natural e Modelos de Linguagem de Larga Escala (LLMs, do inglês *Large Language Models*), a integração de fontes heterogêneas de informação (*e.g.*, conteúdo textual, contexto social) e de conhecimento estruturado em um arcabouço unificado de verificação ainda permanece desafiadora. Ao mesmo tempo, sistemas de AFC continuam a ser avaliados e construídos em torno de rótulos de veracidade, em vez de explicações interpretáveis, robustas e fundamentadas em evidências, necessárias para apoiar checadores de fatos e usuários finais. Esta tese investiga o avanço conjunto da detecção e verificação de desinformação e da explicabilidade ao longo do pipeline de AFC, organizando suas contribuições em detecção de desinformação baseada em classificação por meio de aprendizado de texto e grafos, e em verificação baseada em LLMs fundamentada em conhecimento estruturado e raciocínio. Primeiramente, propomos mu2, um arcabouço multimodal e multilíngue que combina atributos textuais, de metadados e relacionais codificados a partir de postagens em redes sociais para a detecção de desinformação em inglês, espanhol e português, demonstrando como atributos sociais e textuais heterogêneos contribuem para o desempenho da detecção. Em seguida, propomos KG-CRAFT, um arcabouço para verificação de alegações fundamentada em evidências que ancora modelos de linguagem pré-treinados em raciocínio contrastivo baseado em grafos de conhecimento, gerando perguntas contrastivas estruturadas para avaliar a veracidade de alegações e quantificando o ganho atribuível a tal ancoragem em relação a LLMs sem ancoragem. Com base nesses dois arcabouços, abordamos então sua explicabilidade, estendendo mu2 para mu2X, incorporando um Módulo de Explicabilidade da Classificação que produz atribuições *post-hoc* em nível de atributo, avaliado sob protocolos inéditos de confiabilidade e robustez de explicações multimodais. Por fim, estendemos KG-CRAFT com uma etapa de geração de justificativas que interpreta o próprio processo de raciocínio do LLM, avaliando o alinhamento factual e a qualidade das justificativas em linguagem natural

por ele produzidas para seus veredictos. Em conjunto, estas contribuições avançam os dois desafios que motivam esta tese. mu2 e mu2X progridem a integração de fontes heterogêneas de informação em um arcabouço unificado de detecção em múltiplos idiomas, e KG-CRAFT traz o conhecimento estruturado para o raciocínio baseado em LLMs com o objetivo de verificar alegações sob uma perspectiva fundamentada em evidências. No eixo da explicabilidade, mu2X leva os sistemas de AFC para além dos rótulos de veracidade por meio de atribuições em nível de atributo, e KG-CRAFT avança ainda mais ao ancorar a interpretação no próprio raciocínio que produz o veredicto do LLM. Em conjunto, elas levam a área em direção ao objetivo de longo prazo de sistemas de checagem de fatos que sejam, por concepção, precisos, transparentes e confiáveis.

Palavras-chave: Verificação Automatizada de Fatos; Detecção de Desinformação; Aprendizado Multimodal; Grafos de Conhecimento; Modelos de Linguagem de Larga Escala; Inteligência Artificial Explicável.

Abstract

Social media misinformation spreading has outpaced the capacity of manual human-driven fact-checking, motivating the development of automated fact-checking (AFC) systems that can scale with the volume and velocity of online content. However, despite recent advances in natural language processing and Large Language Models (LLMs), the integration of heterogeneous information sources (*e.g.*, textual content, social context) and structured knowledge within a unified verification framework still remains challenging. At the same time, AFC systems continue to be evaluated and built around veracity labels, rather than the interpretable, robust, and evidence-grounded explanations needed to support fact-checkers and end users. This thesis investigates the joint advancement of misinformation detection and verification and explainability across the AFC pipeline, organising its contributions in classification-based misinformation detection based on text and graph learning, and LLM-based verification grounded in structured knowledge and reasoning. First, we propose mu2, a multimodal and multilingual framework that combines textual, metadata, and relational features encoded from social media posts for misinformation detection across English, Spanish, and Portuguese, demonstrating how heterogeneous social and textual features contribute to detection performance. We then propose KG-CRAFT, a framework for evidence-grounded claim verification that grounds pre-trained language models in knowledge graph-based contrastive reasoning, generating structured contrastive questions to assess claim veracity and quantifying the gain attributable to such grounding over ungrounded LLM baselines. Building on these two frameworks, we then address their explainability, extending mu2 into mu2X, augmenting it with a Classification Explainability Module that produces post-hoc, feature-level attributions and is evaluated under novel protocols for the trustworthiness and robustness of multimodal explanations. Finally, we extend KG-CRAFT with a justification generation step that interprets the LLM's own reasoning process, evaluating the factual alignment and quality of the natural language justifications it produces for its verdicts. Together, these contributions advance the two challenges that motivate this thesis. mu2 and mu2X progress the integration of heterogeneous information sources within a unified detection framework across multiple languages, and KG-CRAFT brings structured knowledge into LLM-based reasoning to

verify claims from an evidence-grounded perspective. On the explainability side, mu2X moves AFC beyond veracity labels through feature-level attributions, and KG-CRAFT advances further by grounding interpretation in the reasoning that produces the LLM verdict itself. Together, they advance the field towards the long-term goal of fact-checking systems that are accurate, transparent, and trustworthy by design.

Keywords: Automated Fact-Checking; Misinformation Detection; Multimodal Learning; Knowledge Graphs; Large Language Models; Explainable Artificial Intelligence.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AFC Automated Fact-Checking

BAcc Balanced Accuracy

GAT Graph Attention Network

GNN Graph Neural Network

HSIC Hilbert-Schmidt Independence Criterion

IG Integrated Gradients

KG Knowledge Graph

LIME Local Interpretable Model-agnostic Explanations

LLM Large Language Model

LM Language Model

LoRA Low-Rank Adaptation

MMR Maximal Marginal Relevance

NLP Natural Language Processing

RAG Retrieval-Augmented Generation

RQ Research Question

SLM Small Language Model

XAI Explainable Artificial Intelligence

Contents

1 Introduction

The digital transformation of information ecosystems has fundamentally reshaped how society produces, disseminates, and consumes news, posing unprecedented challenges to information integrity (LEWANDOWSKY; ECKER; COOK, 2017; BAK-COLEMAN et al., 2021; HAIDER; SUNDIN, 2022). The rapid growth of social media platforms as primary channels for news consumption has created an environment in which false and misleading content (Definition 2.2)[1] can spread rapidly and at scale, with documented consequences for democratic processes, public health, and social cohesion (LAZER et al., 2018; VALENZUELA et al., 2019; AÏMEUR; AMRI; BRASSARD, 2023; NEWMAN et al., 2025). Around half of adults in the United States have consistently reported getting news from social media in recent years (48–54% over 2021–2025) (PEW RESEARCH CENTER, 2025), and the shift towards news consumption through social media and video platforms is increasingly pronounced across Latin America, parts of Asia, and Eastern Europe (NEWMAN et al., 2025). On these same platforms, false information propagates faster, farther, and more broadly than accurate information (VOSOUGHI; ROY; ARAL, 2018), particularly during high-stakes events such as elections and public health crises (MAZURCZYK; LEE; VLACHOS, 2024).

Whilst the emergence of professional fact-checking organisations and initiatives such as the International Fact-Checking Network (INTERNATIONAL FACT-CHECKING NETWORK, 2023) has contributed to combating misinformation, the sheer volume and velocity of online content far exceeds the capacity of manual verification (ZHOU; ZAFARANI, 2020; GUO; SCHLICHTKRULL; VLACHOS, 2021). This disparity has motivated the development of automated fact-checking (AFC) systems (Section 2.3), which aim to assist human fact-checkers by automating one or more stages of the verification pipeline (THORNE; VLACHOS, et al., 2018; NAKOV et al., 2021; GUO; SCHLICHTKRULL; VLACHOS, 2021). Recent advances in natural language processing (NLP) and large language models (LLMs) have expanded the capabilities of AFC systems, enabling more advanced approaches to

[1] We use *misinformation* as a general term for false or misleading content, independent of intent (cf. Section 2.1). Because the methods developed here observe only the content and its propagation, not the sender's intent, they do not by construction distinguish *misinformation* from intent-driven *disinformation*.

evidence retrieval, claim verification, and justification generation (AUGENSTEIN et al., 2024; WANG, Yuxia et al., 2024; CHEN; SHU, 2024).

Despite these advances, gaps in the effectiveness of these systems and concerns about transparency persist as fundamental challenges. Claims propagated through social media are shaped not only by their textual content but also by the relational dynamics of the networks through which they spread (SHU; SLIVA, et al., 2017; ALAM et al., 2022). Whilst individual modalities (*e.g.*, text, metadata, social graphs) have each been shown to contribute to detection performance, the systematic *integration of multimodal and structured information sources* within a unified framework remains hard, particularly across multiple languages and cultural contexts (ALAM et al., 2022). Additionally, LLM-based verification systems typically operate on unstructured text, lacking the structured reasoning mechanisms that could ground their assessments in explicit knowledge representations (PAN, S. et al., 2024). A second gap, *explainability and interpretability* (Section 2.4), arises because the increasing complexity of detection and verification models has outpaced the development of methods for explaining their decisions (MINH et al., 2022; SAEED; OMLIN, 2023). Fact-checkers and end users require not only the veracity label but also an understanding of *why* a claim is assessed as true or false, which features contributed to the decision, and how robust those explanations are to perturbation (DOSHI-VELEZ; KIM, 2017; WARREN; SHKLOVSKI; AUGENSTEIN, 2025). Simply labelling content as false without adequate justification can be ineffective or even counterproductive, reinforcing confirmation bias rather than correcting misperceptions (GILPIN et al., 2018; PENNYCOOK; RAND, 2021).

Moreover, these gaps are also connected, as the diversity that drives accurate detection (across modalities for classification and across structured-knowledge relations for verification) is the same diversity that contributes to richer, more interpretable explanations. Multimodal systems can produce richer explanations because each modality contributes different discriminative features about the claim. Structured reasoning over knowledge representations can both improve verification accuracy and support human-readable justifications. Existing systems within each paradigm tend to address modality integration or explainability in isolation, with each property typically pursued without explicit consideration of how design choices in one paradigm transfer to the other.

This thesis addresses these two challenges jointly by proposing methods that leverage multimodal integration and structured reasoning to improve both predictive performance and explainability in automated fact-checking. We develop two framework families,

each with a paired explainability extension: mu2 (LOURENÇO; PAES, 2022a) (and its explainability extension mu2X (LOURENÇO; PAES; WEYDE, 2025)), addressing classification-based misinformation detection over multimodal and multilingual social network content; and KG-CRAFT (LOURENÇO; PAES; WEYDE, et al., 2026) (and its justification generation extension), addressing evidence-based claim verification with LLM-driven knowledge graph-based contrastive reasoning. Across these contributions, we address with two complementary paradigms (classification-based detection and evidence-based verification) and two complementary explanation approaches (post-hoc feature attributions and natural language justifications grounded in structured knowledge), aiming to develop AFC systems that are both more accurate and more interpretable. The remainder of this chapter states the thesis-level hypotheses (Section 1.1) and research questions (Section 1.1.1), maps the chapter structure to the research questions, and enumerates the methodological contributions and supporting publications (Section 1.3).

1.1 Hypotheses

The main hypothesis of this thesis is guided by the following statement:

> *Automated fact-checking systems can simultaneously achieve high veracity-assessment performance and produce explanations of measurably higher quality when the underlying frameworks integrate heterogeneous evidence and reason over its relational structure. Thus, the thesis argues that:*
>
> *(i) the integration of textual content with social network relational structure improves classification-based misinformation detection across diverse linguistic settings, and the resulting multimodal representations support feature-level explanations whose trustworthiness is most stable, with each modality contributing complementary robustness properties under perturbation;*
>
> *(ii) grounding LLM-based claim verification in the structured, relational form of its evidence, and reasoning over that structure through contrastive comparison, improves veracity assessment over approaches that operate on unstructured text alone, whilst the structured reasoning provides the foundations for natural language justifications that approach the quality of human-written reference explanations across multiple explainability desiderata.*

The two clauses (i) and (ii) above are operationalised as four research questions (Research Questions 1 to 4), each derived from one aspect of the corresponding clause and empirically investigated in a subsequent contribution chapter.

1.1.1 Research Questions

This thesis is guided by one general research question, from which four specific research questions are derived, each addressing one aspect of the two hypothesis clauses and investigated in a dedicated contribution chapter.

Research Question (RQ). *How can automated fact-checking systems leverage multimodal integration and structured reasoning to improve veracity assessment whilst producing explanations of measurably higher quality?*

Research Question 1 (RQ1)**.** *How does integrating textual content with social network relational structure affect misinformation detection performance relative to single-modality baselines, and how consistent is this effect across languages?*

Derived hypothesis (clause i) We hypothesise that a jointly encoded multimodal representation of a social network post, combining textual content with social network relational structure, improves on the best single-modality baseline in every language, with the largest improvement in well-resourced languages. Single-modality misinformation detection has typically been built on language-specific text encoders for social media content (NGUYEN; VU; TUAN NGUYEN, 2020; PÉREZ et al., 2022; CARNEIRO et al., 2025) and on graph neural networks (VELIČKOVIĆ et al., 2018; WU et al., 2021), applied to the textual and relational features, respectively. Moreover, multimodal benchmarks combining both signals at scale have only recently emerged, and most evaluations remain English-only (HANGLOO; ARORA, 2022; ALAM et al., 2022; COMITO; CAROPRESE; ZUMPANO, 2023; TUFCHI; YADAV; AHMED, 2023). This research question investigates whether multimodal fusion offers gains across languages or whether the contribution of each modality varies with linguistic resource level. Chapter 3 addresses this question by evaluating a multimodal and multilingual framework on a social media benchmark spanning English, Portuguese, and Spanish.

Research Question 2 (RQ2)**.** *To what extent can structured representations of evidence improve the veracity-assessment capabilities of large language models, compared with reasoning over unstructured evidence?*

Derived hypothesis (clause ii) We hypothesise that grounding LLM-based claim verification in a structured representation of the evidence (instantiated as a knowledge graph extracted from the claim-associated reports), and formulating contrastive questions that probe alternative entity substitutions over it, substantially improve veracity-assessment performance over the baselines. Existing LLM-based fact-checking methods rely on either direct prompting (CHEUNG; LAM, 2023) or retrieval-augmented prompting over unstructured evidence (WANG, B. et al., 2024; XIE et al., 2025), neither of which exploits the relational structure between the entities mentioned in the claim and those in the supporting evidence. This research question investigates whether explicitly extracting the relational structure as a knowledge graph and reasoning over it via contrastive questions yields a measurable improvement in verification accuracy. Chapter 4 addresses this question through a three-phase framework evaluated on `LIAR-RAW` and `RAWFC` with multiple LLM backbones.

Research Question 3 (RQ3)**.** *How does the integration of multimodal features affect feature-level explanation quality across the dimensions of interpretability, trustworthiness, and robustness?*

Derived hypothesis (clause i) We hypothesise that feature-level explanations derived from multimodal post representations are the most stable, whilst each modality contributes a complementary robustness profile under controlled perturbation. Additionally, addressing the interpretability dimension, we examine which feature types most often drive the explanations. Existing explainable misinformation detection systems predominantly rely on content-based (KOU; ZHANG, et al., 2020; SHANG et al., 2022) or comment-based explanations (SHU; CUI, et al., 2019; CUI; SHU, et al., 2019) that explain the decision in terms of the content rather than the specific features that drive it, and their evaluation has been largely confined to interpretability assessments. This research question extends the evaluation to trustworthiness and robustness alongside standard interpretability and asks whether multimodal feature integration is itself a source of explanation quality. Chapter 5 addresses this question by augmenting the multimodal and multilingual framework with a feature-level explainability module and proposing two novel evaluation protocols.

Research Question 4 (RQ4)**.** *How do the choices of evidence representation (structured vs. unstructured) and reasoning formulation (contrastive vs. direct) shape the quality of the natural language justifications produced by LLM-based claim verifiers, when measured across multiple complementary explainability desiderata?*

Derived hypothesis (clause ii) We hypothesise that natural language justifications

produced by an LLM augmented with KG-grounded contrastive reasoning approach the quality of human-written reference explanations on aggregate across multiple explainability desiderata. Existing approaches to LLM-based justification generation evaluate the generated explanations either through small-scale human studies or through reference-based metrics such as ROUGE and BLEU, which combine surface-level lexical overlap with explanatory quality. This research question asks whether structured contrastive reasoning measurably improves the quality of generated justifications, and proposes a desiderata evaluation protocol to assess that improvement across the desiderata. Chapter 6 addresses this question by extending the contrastive reasoning framework with a justification generation step and evaluating it against six desiderata.

1.2 Roadmap

The remainder of this thesis is organised as follows.

Chapter 2 – Theoretical Foundation This chapter presents the foundational concepts and definitions that ground the subsequent contributions. It covers terminology related to misinformation and social networks, the automated fact-checking pipeline, the principles of explainable artificial intelligence (including explanation methods and contrastive reasoning), multimodal data representation, and foundation models.

Chapter 3 – Text and Graph Learning for Misinformation Detection in Social Networks. This chapter introduces mu2, a multimodal and multilingual framework for misinformation detection that integrates textual content, metadata, and relational features encoded from social media post-nodes. It addresses Research Question 1 by focusing on cross-lingual generalisability, evaluating how the combination of social network content and relational dynamics affects detection performance across English, Spanish, and Portuguese on the `MuMiN` benchmark, where multimodal fusion achieves the best mean F1 in each language.

Chapter 4 – Knowledge Graph-based Contrastive Reasoning with LLMs for Misinformation Detection. This chapter presents KG-CRAFT, a framework that leverages LLMs augmented with knowledge graph-based contrastive reasoning for automated claim verification. It addresses Research Question 2 with a focus on the contribution of struc-

tured knowledge grounding to LLM-based verification, investigating whether grounding the reasoning process in knowledge graphs and contrastive question formulation improves the veracity assessment capabilities of pre-trained language models, with state-of-the-art results on `LIAR-RAW` and on `RAWFC`.

Chapter 5 – Explainable Text and Graph Learning for Misinformation Detection in Social Networks. This chapter introduces mu2X, an extension of mu2 augmented with a Classification Explainability Module that provides post-hoc, feature-level explanations for individual classification decisions. It addresses Research Question 3 with a focus on the explanation-quality dimensions beyond standard interpretability by proposing and applying two novel evaluation protocols for the trustworthiness and robustness of multimodal explanations, finding that multimodal explanations are among the most stable under top-K feature selection whilst exhibiting a smoothing position between the most-robust text-based and least-robust graph-based modalities under perturbation.

Chapter 6 – Explainable LLM-based Claim Verification. This chapter extends KG-CRAFT by investigating the natural language interpretation capabilities of the contrastive reasoning framework. It addresses Research Question 4 with a focus on desiderata evaluation of LLM-generated justifications, introducing a six-desideratum LLM-as-a-Judge protocol that scores the quality and factual alignment of generated justifications, finding that KG-CRAFT reaches approximately 87.5% of the human-written reference explanation quality on average and that contrastive question formulation primarily drives explanation quality whilst knowledge graph grounding primarily drives verification accuracy.

Chapter 7 – Conclusion. This chapter summarises the main findings of the thesis, discusses the overarching contributions in light of the research questions, and outlines directions for future work.

1.3 Contributions

1.3.1 Methodological Contributions

The thesis makes the following methodological contributions, each acting on one clause of the main hypothesis (Section 1.1) and addressing one of the research questions (Sec-

tion 1.1.1):

- **mu2** (Chapter 3, Research Question 1): a multimodal and multilingual framework for misinformation detection in social networks that integrates language-specific text encoders with graph attention networks over the post-node representation, evaluated across English, Portuguese, and Spanish on the `MuMiN` benchmark.
- **KG-CRAFT** (Chapter 4, Research Question 2): a three-phase framework for LLM-based claim verification that extracts knowledge graphs from claim-associated reports, formulates contrastive questions over those graphs, and synthesises the answers into a contrastive evidence summary that grounds the verification step.
- **mu2X** (Chapter 5, Research Question 3): an explainability extension of mu2 that augments the classification pipeline with a Classification Explainability Module combining GraphLIME for relational features and Integrated Gradients for textual features, supported by two evaluation protocols for the *trustworthiness* and *robustness* of feature-level explanations that complement standard interpretability assessments.
- **Justification generation extension of KG-CRAFT** (Chapter 6, Research Question 4): an justification generation extension step that consumes the contrastive evidence summary produced by KG-CRAFT's reasoning pipeline, supported by a desideratum evaluation protocol that operationalises six desiderata via LLM-as-a-Judge prompts for the quality and factual alignment of the generated justifications.

1.3.2 Publications

The work proposed in this thesis has been published as conference papers. Chapters 3 to 6 contain existing, improved, or extended results to the following publications:

1. A Modality-level Explainable Framework for Misinformation Checking in Social Networks (LOURENÇO; PAES, 2022a)
   - Main content included in Chapter 3
   - *LatinX in AI Research at the 36th Conference on Neural Information Processing Systems, LXAI-NeurIPS 2022*
2. Exploring Content and Social Connections of Fake News with Explainable Text and Graph Learning (LOURENÇO; PAES; WEYDE, 2025)

- Main content included in Chapter 5
- *35th Brazilian Conference on Intelligent Systems, BRACIS 2025*

3. KG-CRAFT: Knowledge Graph-based Contrastive Reasoning with LLMs for Enhancing Automated Fact-checking (LOURENÇO; PAES; WEYDE, et al., 2026)

- Main content included in Chapter 4
- *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), EACL 2026*

Throughout my doctoral studies, several additional contributions were developed that fall outside the scope of the present thesis, specifically those from past and concurrent positions in the industry. These additional peer-reviewed papers and patents are listed in Appendix C for completeness.

2 Theoretical Foundation

This chapter presents the foundational concepts and definitions that support the work discussed in this thesis. We begin by defining the terminology surrounding misinformation and its dissemination through social networks (Sections 2.1 and 2.2). We then introduce the automated fact-checking pipeline (Section 2.3) and the principles of explainable artificial intelligence (Section 2.4), including explanation methods and contrastive reasoning. Finally, we present the data representation paradigms (Section 2.5) and the foundation models (Section 2.6) employed throughout the subsequent chapters.

2.1 (Fake-, Dis-, Mis-, Mal-) information

Information *noun*

1) news, facts, or knowledge;

2) facts or details about a person, company, product, *etcetera*

Cambridge Dictionary [1].

Information, by its dictionary definition, refers to any piece of content considered as a fact, *i.e.*, in general, a *truthful* statement about a subject. However, when it comes to application in real life, specifically in daily communication, information is usually prone to take a more subjective interpretation of its truthfulness or even reliability. In that sense, different terms were crafted to differentiate the kinds of shared content and their

[1] https://dictionary.cambridge.org/dictionary/english/information

intentions, mostly the pieces of information whose truthfulness or reliability was notably compromised.

Definition 2.1 (Disinformation (WARDLE; DERAKHSHAN, 2017))**.** *Disinformation is any false content that is deliberately created and shared to deceive, manipulate, or otherwise harm its receivers.*

Definition 2.2 (Misinformation (WARDLE; DERAKHSHAN, 2017))**.** *Misinformation is any false or inaccurate content that is shared without the intention to cause harm.*

Definition 2.3 (Malinformation (WARDLE; DERAKHSHAN, 2017))**.** *Malinformation is any genuine content that is shared with the intention to cause harm, often by exposing information meant to remain private.*

To distinguish the different kinds of non truthful information and the intentions behind them, Wardle and Derakhshan (2017) organise these notions along two dimensions, *falseness* and *intent to harm*. As depicted in Figure 1, the structure yields three types: *disinformation* (Definition 2.1), false content shared to cause harm; *misinformation* (Definition 2.2), false content shared without such intent (*e.g.*, inaccurate captions, dates, or statistics, or satire taken at face value); and *malinformation* (Definition 2.3), genuine content repurposed to cause harm. This framework is widely adopted in other works (BRODA; STRÖMBÄCK, 2024) and public-facing guidance (CANADIAN CENTRE FOR CYBER SECURITY, 2024; MUSEUM OF AUSTRALIAN DEMOCRACY AT OLD PARLIAMENT HOUSE, n.d.). Within the framework, the widely used notion of *fake news* is best understood not as a separate category but as an instance of disinformation that additionally mimics the form of legitimate news media (SHU; SLIVA, et al., 2017; ZHOU; ZAFARANI, 2020), with manipulated images and deepfakes among its most prominent manifestations (HANGLOO; ARORA, 2022; TUFCHI; YADAV; AHMED, 2023). Each type can be further categorised into subtypes that vary in intention and harmfulness, from satire and parody to fabricated content, as detailed in other taxonomies (WARDLE, 2017; ALAM et al., 2022). Each type can be illustrated through the 2017 French Presidential election (WARDLE; DERAKHSHAN, 2017): a forged duplicate of a national newspaper falsely claiming a candidate was funded by a foreign state is *disinformation*; the false rumour, unintentionally shared after a terror attack, that a second police officer had died is *misinformation*; and the leak of a candidate's genuine private emails to inflict harm shortly before the vote is *malinformation*. Following common practice in the detection literature (ZHOU; ZAFARANI, 2020; ALAM et al., 2022), we adopt *misinformation* as the operational umbrella term throughout this thesis, denoting false or

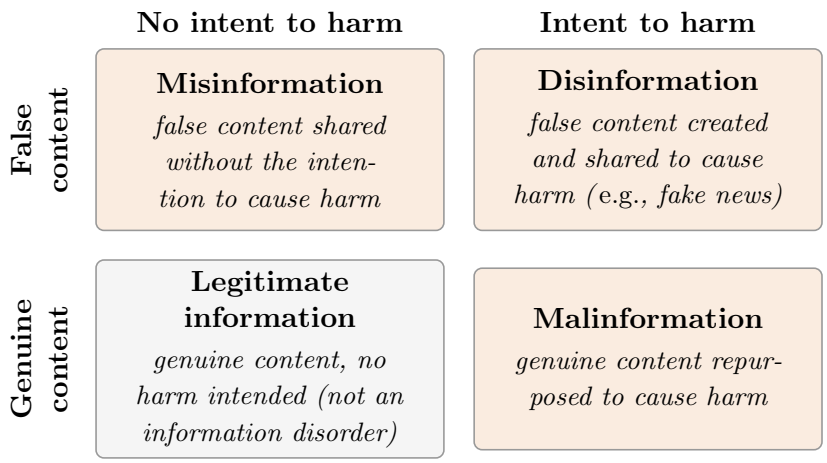


Figure 1: The three types of information disorder along the dimensions of *falseness* and *intent to harm*, following the framework of Wardle and Derakhshan (2017). Fake news is not a separate type but a prototypical instance of disinformation that mimics the form of legitimate news media.

misleading content regardless of the intent behind its dissemination, since the detection methods we develop do not presuppose access to the sender's intent.

These definitions delimit the challenge of assessing the veracity of information core to this thesis. Misinformation is defined in the broad sense aforementioned, where the falseness of the content rather than the sender's intent is what automated fact-checking systems can observe and reason over. This scope is what the detection and verification frameworks of Chapters 3 and 4 target, and whose decisions the explainability extensions of Chapters 5 and 6 render interpretable. Although false and misleading content has accompanied human communication throughout history, the scale, speed, and low cost with which social media enables anyone to create and share content have dramatically amplified its reach and propagation (SHU; SLIVA, et al., 2017; ZHOU; ZAFARANI, 2020). Understanding how this content is structured and disseminated is therefore a prerequisite for detecting it, which motivates the discussion of social media and its content in the next section.

2.2 Social media, networks, and their content

Social media refers to the various online platforms that allow users to create and share content and data with others (SHU; SLIVA, et al., 2017). Social networks, a subset of social media, rely on social interactions and networks of contacts with shared interests. They commonly feature content interaction (such as liking and replying to content), groups, and newsfeeds (ALAM et al., 2022).

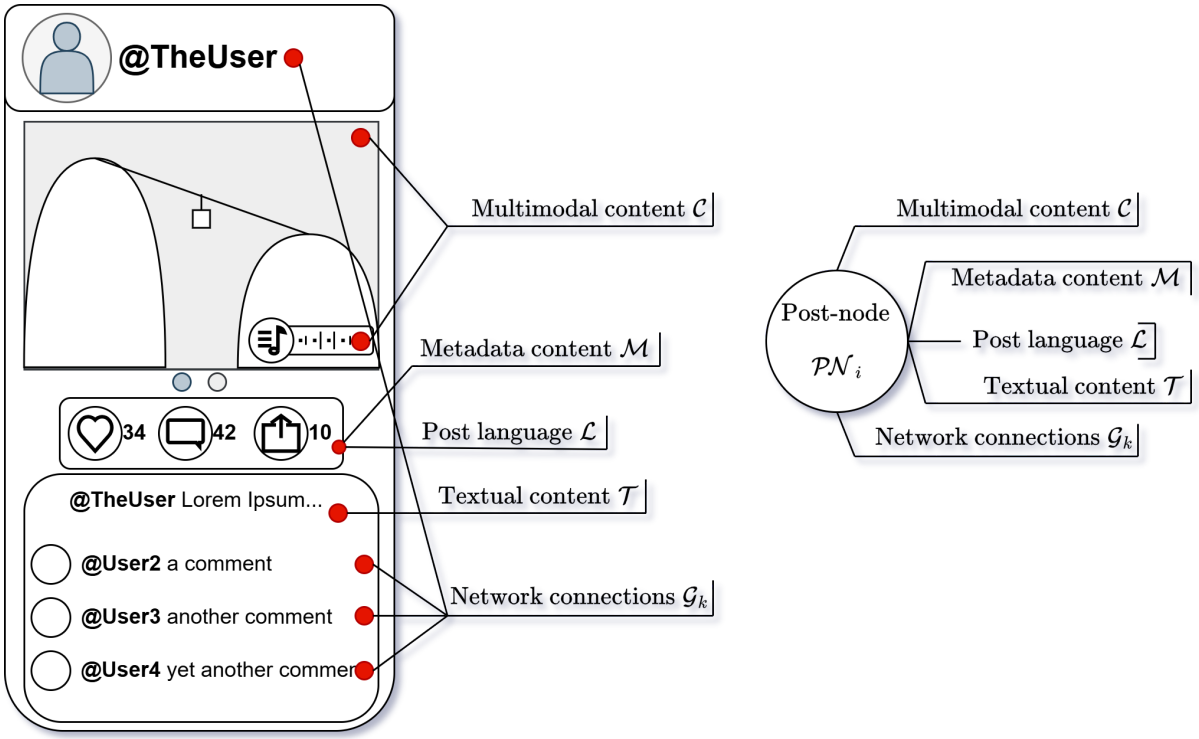


Figure 2: Illustration of a generic social network post (left) and its post-node representation (right).

Social networks connect their users through the pieces of information created and shared by them as digital content, named *post*. As illustrated in Figure 2 (left), a generic post is composed of (but not limited to) textual and multimodal content (*e.g.*, audio, video, and image), metadata (*e.g.*, location, metrics), and social network connections. The format and type of content of a post vary widely depending on the users' intention and the platform characteristics (*e.g.*, Twitter, Instagram, Facebook, LinkedIn, TikTok, *etcetera*).

Definition 2.4 (Post-node)**.** *A post-node is any machine-readable structured representation of a social network's post and its content.*

The post-node (Definition 2.4) is the central unit of analysis for the classification-based detection developed in Chapters 3 and 5, where its textual, metadata, and relational components become the heterogeneous signals that the proposed frameworks integrate and explain.

2.3 Automated Fact-Checking

Definition 2.5 (Automated Fact-Checking)**.** *Automated fact-checking (AFC) is the task of assessing the veracity of a claim using computational methods, typically by retrieving*

and reasoning over relevant evidence.

AFC systems generally follow a multi-stage pipeline comprising four main components: (i) *claim detection*, which identifies check-worthy statements; (ii) *evidence retrieval*, which gathers relevant documents or knowledge from external sources; (iii) *verdict prediction*, which assesses the veracity of the claim given the retrieved evidence; and (iv) *justification generation*, which produces a human-readable explanation supporting the predicted verdict (VLACHOS; RIEDEL, 2014; THORNE; VLACHOS, 2018; GUO; SCHLICHTKRULL; VLACHOS, 2021; WARREN; SHKLOVSKI; AUGENSTEIN, 2025).

In this thesis, we address multiple stages of this pipeline. Chapters 3 and 5 approach verdict prediction as a multimodal classification task over social network content, whilst Chapters 4 and 6 frame it as evidence-based verification using large language models augmented with structured reasoning.

2.4 Explainable Artificial Intelligence

Definition 2.6 (Explainable Artificial Intelligence)**.** *Explainable Artificial Intelligence (XAI) is a set of methods and processes that enable humans to understand and interpret the outputs created by machine learning algorithms.*

The use of XAI emerged with the increasing adoption of non-explainable models (*e.g.*, neural networks, also known as black-box models) in decision-making processes. The main goal is to provide transparency in AI systems (SAEED; OMLIN, 2023), thus facilitating insights into their accuracy, fairness, and the potential biases that might influence their outcomes (MINH et al., 2022). Conversely, many of the concepts that would characterise an explainable AI system are subjective (*e.g.*, fair, responsible, interpretable), primarily due to the social context in which the system is deployed and the different perspectives of the explanations' audience (MILLER, 2019; DOSHI-VELEZ; KIM, 2017; RONG et al., 2024; KIM; MAATHUIS; SENT, 2024; KONG; LIU; ZHU, 2024). The explainability principles introduced in this section serves as foundation of the feature-level attributions of Chapter 5 and the natural language justifications of Chapter 6, the two explanation-oriented frameworks proposed in this thesis.

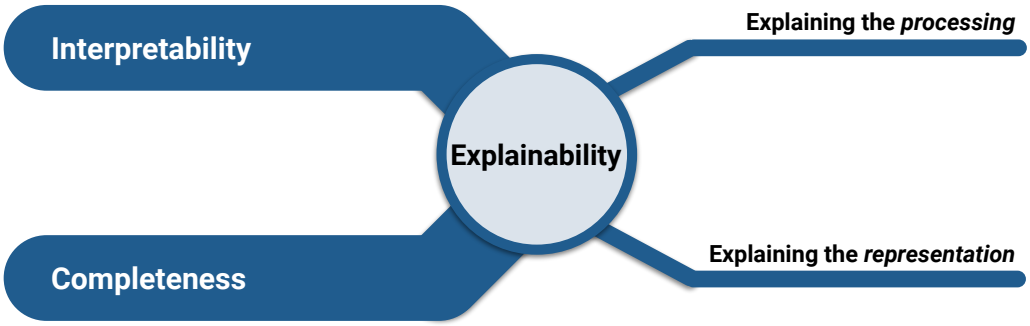


Figure 3: Explainability as a two-dimensional space concept. The goal dimension varies from interpretability to completeness, and the target being the processing or the representation.

2.4.1 Use and perspectives

The need for explainable AI varies according to its use and users' perspective (SAEED; OMLIN, 2023). Five main perspectives can be identified: *regulators* (*e.g.*, judges, politicians), who require compliance with legal standards such as the GDPR; *decision makers* (*e.g.*, domain experts, scientists), who use explanations to validate learned knowledge and guide discovery; *corporate* users, who require both compliant and reliable systems for building trust with customers; *developers* (*e.g.*, ML engineers), who leverage transparency to debug, improve, and de-bias models; and the *general population*, who require trustworthy systems with understandable rationales behind their decisions (SAEED; OMLIN, 2023). In automated fact-checking specifically, the trust between tool developers and professional fact-checkers has been shown to shape whether such systems are adopted in practice (DIERICKX; LINDÉN; OPDAHL, 2023).

2.4.2 Explainability: interpretability and completeness

Explainability can be seen as a two-dimensional space concept, as illustrated in Figure 3. In one dimension, it relates to the explanation's goal: being *interpretable* or *complete*. The other dimension relates to the explanation's target: the *processing* or the *representation*.

Definition 2.7 (Interpretability (GILPIN et al., 2018))**.** *Interpretability aims to describe the internals of a system in a way that is understandable to humans. The success of this goal is tied to the cognition, knowledge, and biases of the user: for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user.*

Definition 2.8 (Completeness (GILPIN et al., 2018)). *Completeness aims to describe the operation of a system in an accurate way. An explanation is more complete when it allows the behaviour of the system to be anticipated in more situations. When explaining a self-contained computer program such as a deep neural network, a perfectly complete explanation can always be given by revealing all the mathematical operations and parameters in the system.*

Definition 2.9 (Explaining the processing (GILPIN et al., 2018)). *Explaining the processing is to describe the manipulation of data within the AI system. It helps to answer questions such as "Why does this particular input lead to that particular output?".*

Definition 2.10 (Explaining the representation (GILPIN et al., 2018)). *Explaining the representation is to describe the internal structure of data in the AI system. It helps to answer questions such as "What information does the AI system contain?".*

The challenge of explainable AI systems is to produce simultaneously interpretable and complete explanations (GILPIN et al., 2018). We argue that they are inversely proportional, given the definition of both concepts, making it impossible to achieve the golden standard in both aspects. The actual challenge is to find the optimal balance between interpretability and completeness for a specific audience.

To illustrate this trade-off, consider a neural classifier that flags a social network post as misinformation. A maximally *complete* explanation would expose every weight, activation, and intermediate computation that led to the verdict. In this setting, the explanation describes the system exactly, yet is unintelligible to anyone but a specialist, and even then only with considerable effort. At the opposite extreme, a single highlighted phrase in the post ("this is the feature that drove the decision") (the kind of feature-attribution explanation studied in Chapter 5) is immediately *interpretable* to a content moderator or end user, but discards almost all of the model's actual behaviour and cannot account for how the verdict would change under different inputs. Neither extreme is satisfactory, where the first is faithful but opaque, the second is accessible but lacks the mathematical grounding. Moreover, the same explanation serves these two audiences unequally, which is why the optimal balance is defined relative to who consumes it.

In this regard, the process of defining and building explainable AI systems has to consider the trade-off between interpretability and completeness. Rather than simplistic explanations (just for the sake of easy interpretation), or exceedingly intricate (outlining every single operation and parameter used), finding a perspective-driven balance between

comprehensiveness (interpretability) and richness (completeness) is paramount for the success of the explanation.

2.4.3 Explanation methods

Explanation methods can be categorised along several axes (MINH et al., 2022; SAEED; OMLIN, 2023; SCHNEIDER, 2024; NGUEAJIO et al., 2025). With respect to phase, *ante-hoc* (*i.e.*, *model-based*) methods build interpretability into the model architecture (*e.g.*, decision trees, attention mechanisms), whilst *post-hoc* methods generate explanations after training by approximating or probing the model's behaviour. With respect to scope, *local* methods explain individual predictions, whilst *global* methods characterise the model's overall behaviour. Finally, methods may be *model-agnostic* (applicable to any classifier) or *model-specific* (designed for a particular architecture).

In this thesis, post-hoc and local explanations are produced by two complementary, model-agnostic methods, applied to the heterogeneous components of mu2X. Ribeiro, Singh, and Guestrin (2016) introduced LIME, which approximates a classifier's local behaviour around an instance x by fitting a sparse linear surrogate to the classifier's predicted scores on perturbed samples in the neighbourhood of x, weighted by a proximity kernel; the surrogate's non-zero coefficients are read as feature attributions. LIME is the foundation for Qiang Huang et al. (2023)'s GraphLIME, which adapts the same locality principle to node classification GNNs by fitting a HSIC-Lasso surrogate (YAMADA et al., 2014) over the target node's N-hop neighbourhood (employed in Chapter 5 for the relational features of mu2X). Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) provides a complementary, gradient-based attribution for the textual features (also in Chapter 5). The mathematical formulations and implementation details of these methods are presented in Chapter 5.

2.4.4 Contrastive reasoning and explanations

Contrastive explanations address questions of the form "Why P rather than Q?", where P is the *fact* (the observed outcome) and Q is the *foil* (an expected alternative that did not occur) (LIPTON, P., 1990; MILLER, 2019). The foil is distinct from the counterfactual of causal analysis, which denotes a hypothetical cause (the case in which a candidate cause C does not occur), whereas a foil is a hypothetical outcome, *i.e.*, an alternative event Q that the explainee expected instead of P (MILLER, 2019). Following Peter Lipton (1990),

Miller (2019) argues that essentially all *why*-questions are contrastive (even when the foil is left implicit) and that contrastive answers are both easier to produce and easier to evaluate than complete causal accounts. Consequently, a contrastive explanation needs only to isolate the differences that make P actual rather than Q, since the fact and the foil typically share most of their causal history.

As an example, consider a fact-checking system that labels the claim *"vaccine V causes condition C"* as *false*. A non-contrastive question, *e.g.*, *Why is the claim false?*, asks for a complete causal account. To answer it fully, the system would have to list every counter-evidence, lack of corroboration, and the reasoning steps behind the verdict, which are elements that a real system cannot produce. The contrastive form, *Why is the claim false rather than true?*, is more tractable. The system needs only cite the differences between the false-supporting and true-supporting evidence, such as the absence of any peer-reviewed study linking V to C alongside the presence of regulatory rebuttals. The fact and the foil share the same history, *i.e.*, the same claim and the same evidence retrieval process, and the explanation isolates only what drove the verdict. This is also the structure of evidence a human fact-checker is trained to produce.

In XAI, this paradigm has predominantly been operationalised *post-hoc*, through counterfactual methods that identify the minimal set of feature changes that would alter a model's prediction (WACHTER; MITTELSTADT; RUSSELL, 2017; JACOVI et al., 2021; STEPIN et al., 2021). Aryal and Keane (2023) extend this line by formalising *semi-factual* ("even if") explanations alongside the more familiar counterfactual ("if only") form, providing benchmarks and desiderata for both. The contrastive and counterfactual paradigms are related but distinct. Contrastive explanations (LIPTON, P., 1990; MILLER, 2019) ask *Why P rather than Q?* with Q a hypothesised outcome, whereas counterfactual explanations in the XAI sense ask which input perturbation would change the verdict. Chapter 4 builds on the former.

A consequential property of contrastive explanations is that the act of producing them is itself a structured reasoning trace. Posing "Why P rather than Q?" forces a system to compose the decision procedure through three steps: *(i)* identify the foil Q; *(ii)* align the causal histories of the fact and the foil; and *(iii)* isolate the differences that account for P. In this thesis, we exploit this interchangeability: rather than treating the contrastive question as a post-hoc probe of an already-trained model (the dominant XAI paradigm), Chapter 4 formulates contrastive questions as the *primary inference loop* of the verification framework, so that the steps a human would request to *understand* the verdict are the

same steps the system performs to *reach* it. This collapses the conventional separation between reasoning and explanation, and is the conceptual basis for the explanation-quality gains evaluated in Chapter 6.

Chapter 4 formulates contrastive questions as the primary inference loop of the verification framework, rather than as a *post-hoc* probe of a trained model. Consequently, the steps a human would request to understand the verdict are the same steps the system performs to reach it. Reasoning and explanation are no longer separate stages but interchangeable, which grounds the explanation quality gains evaluated in Chapter 6.

2.5 Multimodal Data Representation

Multimodal learning refers to the integration of information from multiple heterogeneous data sources (modalities) to improve model performance and provide richer representations (BALTRUŠAITIS; AHUJA; MORENCY, 2019; ALAM et al., 2022). In the context of misinformation detection, the relevant modalities include textual content (the claim or post text), metadata (engagement counts and post-level statistics), and relational structure (social network graphs, knowledge graphs). Strategies for combining these modalities range from *early fusion* (concatenating feature vectors before classification) to *late fusion* (combining modality-specific predictions) and *hybrid* approaches that mix the two (ATREY et al., 2010; BALTRUŠAITIS; AHUJA; MORENCY, 2019).

The relational modality plays an important role in this thesis, as it carries information that the textual and metadata modalities cannot express on their own. We utilise the *graphs* paradigm, which encode social networks' propagation and interaction structure surrounding a post, as described in Section 2.5.1, and *knowledge graphs* paradigm, which encode the factual structure of the claim and associated reports that asserts or contradicts them claim, as seen in Section 2.5.2. Graph neural networks provide the architectural means to learn over the former, and structured triple representations provide the symbolic means to reason over the latter, the two subsections that follow introduce each in turn. The classification frameworks of Chapters 3 and 5 fuse social network graph features with text and metadata via early fusion, and the verification framework of Chapter 4 grounds structured reasoning in claim-specific knowledge graphs using language models.

2.5.1 Graph neural networks

Graph neural networks (GNNs) emerged from the observation that the dominant deep learning architectures, *e.g.*, convolutional networks for grids of pixels and recurrent networks for sequences of tokens, presuppose a regular and ordered domain that graph-structured data does not have (WU et al., 2021). Graph-encoded structures, such as social networks, maps, or molecules, have structural properties (*e.g.*, irregular topology, permutation-invariant neighbourhoods, and a variable number of connections per node), which are unfitted to the inductive biases of CNNs or RNNs. Scarselli et al. (2009) introduced the first general formulation of a neural network operating directly on graphs through an iterative state-propagation procedure. The graph convolutional network (KIPF; WELLING, 2017) bridged the two and provided the practical breakthrough that made GNNs scalable, after which Gilmer et al. (2017) unified the disparate variants under a single *message-passing* abstraction.

Under this abstraction, a single GNN layer updates the hidden state $\mathbf{h}_v^{(l)}$ of each node v through three operations. First, it computes a *message* $\mathbf{m}_{u\to v}^{(l)} = \mathrm{MSG}(\mathbf{h}_u^{(l)}, \mathbf{h}_v^{(l)}, e_{uv})$ for every neighbour $u \in \mathcal{N}(v)$, optionally utilising edge features e_{uv}. Second, a permutation-invariant *aggregation* operator (*e.g.*, sum, mean, or max) combines these incoming messages, yielding $\mathbf{a}_v^{(l)} = \mathrm{AGG}(\{\mathbf{m}_{u\to v}^{(l)} : u \in \mathcal{N}(v)\})$. Finally, an *update* function computes the node's next state using the aggregated signal and its current state, $\mathbf{h}_v^{(l+1)} = \mathrm{UPDATE}(\mathbf{h}_v^{(l)}, \mathbf{a}_v^{(l)})$. Stacking L layers allows nodes to capture structural and semantic information from their L-hop neighbourhood. This message-passing step is illustrated in Figure 4(a) and we refer to Wu et al. (2021) for a broader survey of GNN architectures.

Graph Attention Networks (GATs) (VELIČKOVIĆ et al., 2018) specialise this aggregation step. Instead of using static, degree-normalised weights like standard graph convolutions, GATs compute a learned, feature-dependent attention coefficient α_{vu}, normalised via a softmax over $\mathcal{N}(v)$. In GATs, the aggregated signal becomes a weighted sum, $\mathbf{a}_v^{(l)} = \sum_{u\in\mathcal{N}(v)} \alpha_{vu}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)}$. This formulation allows the network to end-to-end weighted more informative neighbours whilst reducing noise from irrelevant ones, an inductive bias that benefits modelling misinformation. A social network graph in which a claim is the central node connected to a set of reposts, quotes, and replies is often noisy, with neighbours of different types carrying unequal weights, for instance, a quoting post-node that engages directly with the claim is more informative than one that simply propagates it. Figure 4(b) demonstrates how a GAT dynamically assigns larger attention weights α_{vu} to the more informative neighbours. Consequently, this thesis relies on GATs rather than feature-

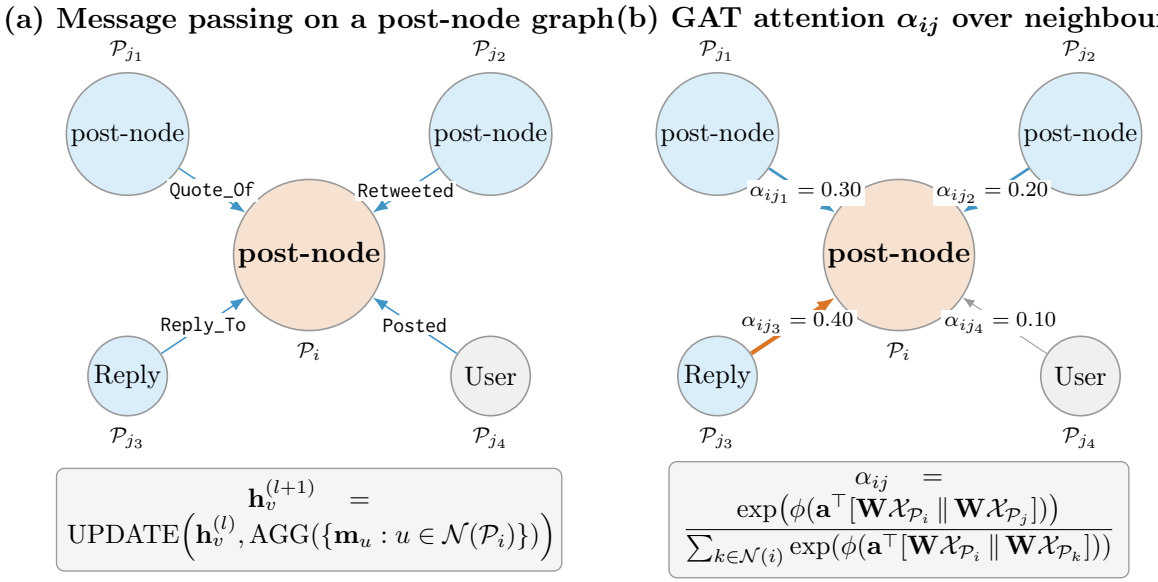


Figure 4: One round of message passing (left) and GAT attention (right) on a post-node graph (Definition 2.4). The central post-node $\mathcal{P}_i$, has four heterogeneous neighbours drawn from the relations used in Chapter 3 (`Quote_Of`, `Retweeted`, `Reply_To`, and `Posted`). (a) aggregates the neighbour messages over the neighbourhood $\mathcal{N}(\mathcal{P}_i)$ and combines them with the node's previous state to produce $\mathbf{h}_v^{(l+1)}$. (b) replaces uniform aggregation with learned attention coefficients α_{ij}. The non-linear activation $\phi(\cdot)$ is the LeakyReLU standard to the GAT formulation (VELIČKOVIĆ et al., 2018).

agnostic convolutions, enabling the model to prioritise relevant content over raw graph topology.

In Chapter 3, mu2 fuses content (text and metadata) with relational features by passing a post-node graph (Definition 2.4) through a GAT classifier trained on the multilingual MuMiN benchmark (NIELSEN; MCCONVILLE, 2022), where nodes represent claims, tweets, replies, and users, and edges encode the propagation, authorship, and topical structure (`Posted`, `Mentions`, `Retweeted`, `Quote_Of`, `Reply_To`, and `Discusses` relations). Chapter 5 extends the same backbone in mu2X with GraphLIME explanations, leveraging the locality principle introduced in Section 2.4.3 to attribute predictions to a small set of influential neighbours. Across both chapters, the relational signal recovered by message passing complements text-only baselines that ignore propagation structure.

2.5.2 Knowledge graphs

A knowledge graph (KG) is a structured representation of factual knowledge as a set of triples (e_h, r, e_t), where $e_h, e_t \in \mathcal{E}$ are entities and $r \in \mathcal{R}$ is a relation connecting them (HOGAN et al., 2021; JI et al., 2022). A claim such as "COVID-19 is caused by

SARS-CoV-2" is then expressed by a triple of the form (`COVID-19`, `causedBy`, `SARS-CoV-2`) that an automated system can store, query, and reason over. In contrast to the social network graphs of Section 2.5.1, where nodes typically represent users or posts and edges encode interactions, a KG encodes *semantic* relationships between real-world entities, and large general-purpose KGs such as Wikidata (VRANDEČIĆ; KRÖTZSCH, 2014) and DBpedia (LEHMANN et al., 2015) provide millions of such triples to ground downstream tasks.

KGs can be constructed either by curating an ontology and populating it manually (the route Wikidata takes) or by extracting triples automatically from unstructured text, a task that decomposes into named-entity recognition, entity linking, and relation extraction (JI et al., 2022). The latter route has been transformed by large language models, which can identify entities and relations in zero- or few-shot methods without task-specific training (ZHU et al., 2024; ZHANG; SOH, 2024). An addition line of work integrates KGs with language models at inference time, retrieving relevant triples to ground generation in verifiable structured facts (PAN, S. et al., 2024; PENG et al., 2025).

This thesis adopts the LLM-based extraction route in Chapter 4, where claim-specific KGs are constructed at inference time from each claim's evidence reports rather than relying on a pre-existing encyclopaedic KG, ensuring that the structure made available for verification is exactly the structure the cited evidence supports. The contrastive reasoning of Section 2.4.4 is then carried out over the resulting KG, and the prompting techniques used to drive the extraction itself are introduced in Section 2.6.

2.6 Large Language Models

Large language models (LLMs) are a particular instance of foundation models, that is, large-scale neural networks pre-trained on broad data corpora at a scale that confers broad task generality and that are subsequently adapted to downstream tasks (BOMMASANI; LIANG, et al., 2021). The dominant architecture underlying modern LLMs is the Transformer (VASWANI et al., 2017), whose central mechanism is scaled dot-product attention. In this mechanism, each token in a sequence is mapped to query, key, and value vectors, and a weighted sum over values is computed using attention weights derived from the dot products of queries and keys, normalised by a softmax and scaled by $\sqrt{d_k}$. Stacking many such attention layers, interleaved with feed-forward sublayers, yields a model that can capture long-range dependencies in sequential data without the recurrence of earlier

sequence models.

Two pre-training objectives partition the LLM landscape relevant to this thesis. *Encoder* models such as BERT (DEVLIN et al., 2019) are trained with masked language modelling, where a fraction of tokens in the input are hidden and the model learns to recover them from bidirectional context. The resulting contextualised representations are well suited to classification tasks, and Chapters 3 and 5 use BERT-family encoders to embed the textual content of social-media post-nodes. *Decoder* models such as GPT-3 (BROWN et al., 2020), Llama 3 (GRATTAFIORI et al., 2024), and Claude 3 (ANTHROPIC, 2024) are trained with causal language modelling, predicting each token from those that precede it, and produce text autoregressively. These generative models are used in Chapters 4 and 6 for evidence-based veracity verification and natural language justification production.

A defining property of generative LLMs is that they can be adapted to new tasks through *prompting* alone, without parameter updates (BROWN et al., 2020; DONG et al., 2024). *Zero-shot* prompts provide only a task description, whereas *few-shot* prompts include a small number of input and output examples that the model imitates as part of its in-context learning (DONG et al., 2024). More structured prompting strategies elicit explicit intermediate reasoning steps before the final answer, the most influential being chain-of-thought prompting (WEI, Jason et al., 2022), with broader reasoning paradigms surveyed by Huang and Chang (2023). Chapter 4 adopts phased few-shot prompting for knowledge graph extraction, decomposing entity identification, category classification, and relation extraction into a sequence of structured prompts. Subsequent phases (contrastive question generation, contrastive question answering, and answer summarisation) use further structured prompts grounded in the extracted graph.

Retrieval-augmented generation represents another line of work capable of mitigating hallucination by conditioning an LLM's output on external documents retrieved at inference time (LEWIS et al., 2020; GAO, Y. et al., 2023). Furthermore, a parallel paradigm grounds LLM reasoning in structured knowledge graphs (Section 2.5.2) (PAN, S. et al., 2024; PENG et al., 2025). Chapter 4 follows the latter, extracting a claim-specific knowledge graph from the claim's associated reports and using it to drive contrastive reasoning, and Chapter 6 evaluates the quality of the resulting natural language justifications. Combined with the encoder-based classifier of Chapters 3 and 5, these LLM-based components complete the methodology used in the remainder of this thesis.

Part I

Misinformation Detection

3 Text and Graph Learning for Misinformation Detection in Social Networks

This chapter presents mu2, a multimodal and multilingual framework for misinformation detection in social networks that integrates textual and relational features encoded from social media post-nodes. The explainability dimension of mu2, which elucidates the features driving the classification decisions, is addressed in Chapter 5.

The main content is an extended version of the paper "A Modality-level Explainable Framework for Misinformation Checking in Social Networks" published in the *LatinX in AI Research at the 36th Conference on Neural Information Processing Systems, LXAI-NeurIPS 2022* (LOURENÇO; PAES, 2022a).

3.1 Introduction and Background

This chapter addresses the task of detecting misinformation (Definition 2.2) in social networks. The dissemination of such content across social media platforms has been observed to impact democratic processes (VALENZUELA et al., 2019), erode public trust in institutions (TEWKSBURY; RITTENBERG, 2012), and, in the case of health-related misinformation, directly endanger lives (MAZURCZYK; LEE; VLACHOS, 2024). With social media serving as a major source of news consumption worldwide (NEWMAN et al., 2025), there is a growing need for automated methods for misinformation detection and mitigation; however, this remains a challenging problem (SHU; SLIVA, et al., 2017; AUGENSTEIN et al., 2024).

We focus on multimodal misinformation detection methods that assess the veracity of social media content by jointly leveraging textual information and the relational structure of the underlying network, as formalised by the post-node representation in Section 2.2. State-of-the-art models for this task are predominantly driven by advances in pre-trained language models (Section 2.6) (NGUYEN; VU; TUAN NGUYEN, 2020; PÉREZ et al., 2022; CARNEIRO et al., 2025) and, separately, graph neural networks (Section 2.5.1) (VELIČKOVIĆ et al.,

2018), which encode, respectively, the semantic content and the structural relationships between users, posts, and metadata. However, most methods rely on one of these encoder families in isolation. Text-only models misclassify content that mimics the linguistic style of legitimate news at the surface level whilst missing the propagation information carried by who shares it and how (SHU; SLIVA, et al., 2017; ZHOU; ZAFARANI, 2020). Conversely, graph-only models can detect suspicious cascades but cannot distinguish a viral fact-check from a viral falsehood without recourse to the underlying content (BIAN et al., 2020; SHU; MAHUDESWARAN, et al., 2020). Furthermore, the few approaches that do combine modalities are predominantly evaluated on English-only benchmarks (HANGLOO; ARORA, 2022; ALAM et al., 2022; COMITO; CAROPRESE; ZUMPANO, 2023; TUFCHI; YADAV; AHMED, 2023), leaving open whether their conclusions transfer to languages with different network practices, encoder ecosystems, and fact-checking infrastructures. We argue that misinformation detection particularly benefits from integrating textual and graph-based features within a multilingual framework, enabling a systematic understanding of how content and relational interact and trade off across languages. Because mu2 draws on propagation and engagement features, it operates in the *late detection* setting, assessing content after it has begun to circulate.

In this chapter, we empirically analyse the impact of multimodal feature integration. We combine language-specific text encoders with graph attention networks to detect misinformation across multiple languages. We evaluate mu2 on the `MuMiN` dataset (NIELSEN; MCCONVILLE, 2022), a large-scale multilingual benchmark comprising 2,183 fact-checked claims and over seven million tweets spanning English, Portuguese, and Spanish, using language-specific encoders (NGUYEN; VU; TUAN NGUYEN, 2020; PÉREZ et al., 2022; CARNEIRO et al., 2025) and graph attention networks (VELIČKOVIĆ et al., 2018). Our experiments are designed to investigate the relative contribution of textual and graph-based features, both individually and in combination, across diverse linguistic contexts. This chapter addresses Research Question 1, which asks how integrating textual content with social network relational structure affects detection performance relative to single-modality baselines, and how consistent this effect is across languages. The main contributions of this chapter are as follows:

- We propose mu2, a multimodal and multilingual framework that integrates language-specific text encoders with graph attention networks over social network structure for misinformation detection;
- We systematically evaluate the relative contribution of textual and graph-based

features across three languages, demonstrating that multimodal fusion consistently achieves the best performance; and

- We provide empirical evidence that text-based features are more robust than graph-based features alone in low-resource and sparse graph settings, offering practical insights for multilingual deployment.

In the following sections, we first survey the related work on multimodal misinformation detection and the datasets designed to support this task (Section 3.2). We then present the mu2 framework, detailing the post-node encoding, relational representation, and classification components (Section 3.3). Subsequently, we describe the experimental setup (Section 3.4), report and analyse the results (Section 3.5), and conclude with a discussion of findings and future directions (Section 3.6).

3.2 Related Work

This section surveys the literature most relevant to the contributions of this chapter across multimodal misinformation detection and the datasets designed to support this task. The former situates mu2 within the current state of the art, whilst the latter motivates the selection of `MuMiN` (NIELSEN; MCCONVILLE, 2022) as the primary benchmark for our experimental evaluation.

3.2.1 Multimodal Misinformation Detection

The current landscape of automated misinformation classification is characterised by a significant emphasis on methodologies that prioritise single-modality content, often at the expense of overlooking the relationships between different data types. For instance, Linmei Hu et al. (2021) leverages graph neural networks supported by external knowledge and Pawan Kumar Verma et al. (2021) rely on word embeddings fused with diverse linguistic features (*e.g.*, readability indices, stylistic and psycho-linguistic attributes among other features). Other work, such as in (KAWINTIRANON; SINGH, 2023), utilises reinforcement learning for textual fact-checking, or focus exclusively on encoding visual information to identify manipulated content (QI et al., 2019; SINGH; SHARMA, 2021). Although these works propose valuable solutions for addressing misinformation within the social media ecosystem, they remain constrained by their reliance on individual modalities (*e.g.*, text, and images) in isolation. Consequently, by analysing text or images as independent signals,

these approaches overlook the inherent benefits of reasoning over the complex, cross-modal information embedded within the naturally multimodal structure of social media content.

Motivated by this inherent complexity, various novel techniques have emerged to leverage the diverse modalities in social media content. The proliferation of these methodologies has been extensively documented in several recent surveys of multimodal misinformation detection (HANGLOO; ARORA, 2022; ALAM et al., 2022; COMITO; CAROPRESE; ZUMPANO, 2023; TUFCHI; YADAV; AHMED, 2023; ABDALI; SHAHAM; KRISHNAMACHARI, 2024; LI, X. et al., 2025), which systematically categorise the literature based on distinct criteria, including modality types, underlying tasks, and data sources. Specifically, these reviews provide a structured characterisation of the field by grouping existing works according to their methodological frameworks and the specific languages addressed. Building upon this foundation, we situate the contribution of this chapter within the current landscape by discussing, in the following, the most relevant approaches, focusing specifically on misinformation detection on social networks from a multimodal perspective and linguistic coverage – aspects that remain central to our proposed scope.

Social media content is primarily characterised by textual data, which encapsulate the core of the communicated message, and user interaction signals (*e.g.*, replies, reactions, and re-posts), that define the structural relationship between a post and its wider network; hence, text-based and graph-based features (also referred to as network features) serve as representations of both the communicated content and the structural relationships between users, posts, and metadata. Consequently, the integration of both sets of features has emerged as a primary strategy for encoding both explicit semantic content and the latent network connections (HAMED; AB AZIZ; YAAKUB, 2023). Early methodologies in this domain frequently utilised propagation networks to capture these dynamics. For instance, Shu, Wang, and Liu (2019) propose the fusion of textual and network features to automate misinformation detection. Their findings suggested that structural network information provides more indicative cues for identifying fake news than the textual message itself. Furthermore, (SHU; MAHUDESWARAN, et al., 2020) addressed the same task by utilising a hierarchical propagation network-based method, with experimental results demonstrating that graph-encoded information is more informative for misinformation identification than purely textual representations.

The aforementioned work limits its scope to English-only content, which regionalises the analysis and limits the generalisability of findings regarding the prominence of graph features over textual message content. Our method, described in Section 3.3, leverages the

combination of both text-based and graph-based features to perform the misinformation detection task; specifically, we conduct our analysis on a multilingual dataset, facilitating a clearer understanding of how feature impact varies across diverse languages (Sections 3.3.3 and 3.5).

Another frequent characterisation of social media content is the combination of textual and visual modalities, such as the use of images to supplement written narratives. Due to the intuitive nature of visual information, posts containing images are consumed more easily and spread more rapidly, leading to increased user interaction (ALAM et al., 2022). Therefore, studies aim to understand and relate these two distinct modalities to improve the performance of automated misinformation detection methods. For instance, (QIAN; WANG, et al., 2021) propose HMCAN, an attention-based multimodal network for fake news detection, which jointly models multimodal context information and the hierarchical semantics of text. Specific focus on "fauxtography" [1] was introduced by (WANG, Y. et al., 2021), targeting misinformation grounded in modified or miscaptioned images intended to mislead the audience. Building upon structured representations, (QIAN; HU, et al., 2021) developed KMAGCN, utilising background knowledge triplets alongside text and image features within an adaptive graph convolutional network to detect fake news. Another relevant approach is SAFE (ZHOU; WU; ZAFARANI, 2020), which performs misinformation detection via joint representation learning of textual and visual features. Specifically, the method utilises neural networks for feature extraction and a cross-modal similarity mechanism to identify potential mismatches between the two modalities. Liu, Wang, and Li (2023) approach multimodal misinformation detection as a logic-reasoning problem over text and image, jointly improving predictive performance and the interpretability of the decision, whilst Cui and Shang (2024) combine cross-modal interaction with graph contrastive learning over text and image to detect fake news. Experimental results for these methods confirm that leveraging both images and text significantly enhances predictive performance for the classification task.

Despite methodological variations, these works prioritise the growing use of images in misinformation without incorporating network-based features within their scope. We note that KMAGCN (QIAN; HU, et al., 2021) utilises relational triplets derived from external knowledge bases; however, this reliance on external information introduces significant scalability constraints associated with the time-intensive nature of data sourcing and validation, which may limit practical utility in real-world deployment. Moreover, consistent

[1] Misleading presentation of images for propagandistic or otherwise ulterior purposes, involving staging, deceptive modification, or the addition or omission of significant context.

with the observations made previously in this section, these methods are restricted to English-only content, underscoring the need for further investigation into multilingual applications.

More recently, attention has turned to large language models (LLMs) as a component of misinformation detection pipelines; Beizhe Hu et al. (2024), for instance, examine the role LLMs can play in fake news detection, finding them more effective as advisors that complement smaller task-specific models than as standalone classifiers. This direction is orthogonal to the multimodal, multilingual feature integration studied in this chapter and is taken up for evidence-based claim verification in Chapter 4.

3.2.2 Multimodal Misinformation Datasets

In this section, we subsequently detail the specific features of selected existing datasets designed for the misinformation detection task. These datasets are characterised by significant variations in linguistic coverage, thematic topics, and the types of available content.

Fakeddit (NAKAMURA; LEVY; WANG, 2020) Derived from Reddit[2], this dataset includes over 1 million posts produced by over 300,000 users collected from 22 highly engaged, popular, and diverse subreddits between 19 March 2008 to 24 October 2019. `Fakeddit` is multimodal, including the textual and visual content of posts, as well as relational aspects such as comments and metadata, with 64% of posts containing both textual and visual information. The dataset is categorised with labels, covering classifications such as true, satire/parody, misleading content, manipulated content, false link and imposter content.

ReCOVery (ZHOU; MULAY, et al., 2020) Comprising over 2,000 news articles from a diverse range of 60 news sources, `ReCOVery` is a multimodal dataset (2,029 news articles, mostly text-image pairs) for exploring news and misinformation spread on social media. Notably, it exhibits a real-world-like class imbalance (2:1 real-to-fake ratio) and captures user activity (78,659 sharing real news, 17,323 sharing fake news), suggesting potential user overlap in dissemination. News source credibility is validated by NewsGuard[3] and MBFC[4].

[2] https://www.reddit.com/
[3] https://www.newsguardtech.com/
[4] https://mediabiasfactcheck.com/

MuMiN (NIELSEN; MCCONVILLE, 2022) `MuMiN` is a relational and multimodal dataset extracted from Twitter[5]. Specifically, `MuMiN` links tweets covering multiple topics and languages with fact-checked claims, including text, metadata, and visual content from the tweets. It has posts written in 70 different languages, spanning over 21 million tweets and nearly 13,000 fact-checked claims from 115 fact-checking organisations. At the time of mu2's development, `MuMiN` was, to our knowledge, the largest published dataset of its kind to support research on automatic misinformation detection in social media.

FactDrill (SINGHAL; SHAH; KUMARAGURU, 2022) Focusing on providing resources for misinformation detection studies in India, `FactDrill` is a dataset that contains over 22,000 fact-checked social media content from 2013 to 2020 that incorporates content in 13 different languages spoken in the country. Furthermore, it is characterised by providing 14 different attributes divided into five feature sets: metadata, textual, media, social, and event. The metadata features include various details of the post (*e.g.*, source name, date of publication, article link). The textual features include the actual text content of the social media post being fact-checked. The media features encompass information about any multimodal content (*e.g.*, images, videos, and audio) associated with the post. The social features include details such as user engagement metrics (*e.g.*, shares, likes) or information about the fact-checking organization responsible for the evaluation. Finally, the event features comprise information about the real-world event or topic the social media post refers to.

MR2 (HU, X. et al., 2023) `MR2` is a multimodal collection of 14,700 English and Chinese social media posts that are labelled as rumour, non-rumour, or unverified. Each post includes text, image, and metadata (*e.g.*, publication date, number of likes and reposts) retrieved from Twitter and Weibo[6]. In addition to the content of the post itself, `MR2` also provides retrieved evidence from the internet, including textual descriptions of relevant images and other social media posts that mention the image.

Mocheg (YAO et al., 2023) This dataset is designed to address the limitations of existing fact-checking datasets by incorporating multimodality (text and image), human-annotated labels (supported, refuted, or not enough information) and human-written explanations. It consists of 15,601 claims collected from the fact-checking websites (*e.g.*, Snopes[7] and

[5] https://twitter.com/ [6] https://weibo.com/ [7] https://www.snopes.com/

PolitiFact[8]) with 33,880 textual paragraphs and 12,112 images in total as evidence.

For our experimental setup (detailed in Section 3.4), we utilise the `MuMiN` dataset (NIELSEN; MCCONVILLE, 2022) to carry out the evaluations of our proposed method. This selection is motivated by the inclusion of multilingual and multi-topical texts, as well as the network data required to contextualise these signals within the social media environment.

3.3 Method

This section introduces mu2, a multimodal and multilingual framework for detecting misinformation in social media posts across multiple modalities, languages, and topics. We initiate this discussion by defining the problem of classifying misinformation in social media posts within a multimodal context. Subsequently, we provide a high-level architectural overview of the individual components of mu2, followed by a detailed characterisation of their technical implementation.

3.3.1 Problem Statement

We formalise the misinformation classification task as a supervised learning problem over post-nodes. A post-node $\mathcal{P}_i$ is the tuple

$$\mathcal{P}_i = (\mathcal{T}_i, \mathcal{G}_k^{(i)}, \mathcal{M}_i, \mathcal{L}_i), \tag{3.1}$$

where $\mathcal{T}_i$ represents the textual content of the post; $\mathcal{G}_k^{(i)} = (V_i, E_i)$ is the local k-hop neighbourhood of $\mathcal{P}_i$ in the underlying graph, with V_i the set of post-nodes within k hops of $\mathcal{P}_i$ and $E_i \subseteq V_i \times V_i \times \mathcal{R}_{\text{social}}$ the typed edges drawn from the relation set $\mathcal{R}_{\text{social}}$ $\mathcal{M}_i \in \mathbb{R}^{d_\mathcal{M}}$ is the engagement feature vector, encoding post-node's metadata such as reply, quote, and repost counts; and $\mathcal{L}_i \in \mathbb{L}$ is the language of the post, drawn from a finite set of supported languages $\mathbb{L}$.

The language $\mathcal{L}_i$ is treated as a separate component rather than as a metadata field because it is consumed for routing, selecting which language-specific text encoder is applied to $\mathcal{T}_i$, rather than as a numeric feature encoded into $\mathcal{M}_i$ (Section 3.3.3).

Given a labelled dataset $\mathcal{D} = \{(\mathcal{P}_i, y_i)\}_{i=1}^N$ where $y_i = f^\star(\mathcal{P}_i) \in \{0, 1\}$ is the true labelling function (with 0 denoting *misinformation* and 1 denoting *fact*), the objective is to train a classifier $f_\theta : \mathcal{P} \to [0, 1]$ such that the predicted probability $\hat{y}_i = f_\theta(\mathcal{P}_i)$ minimises

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

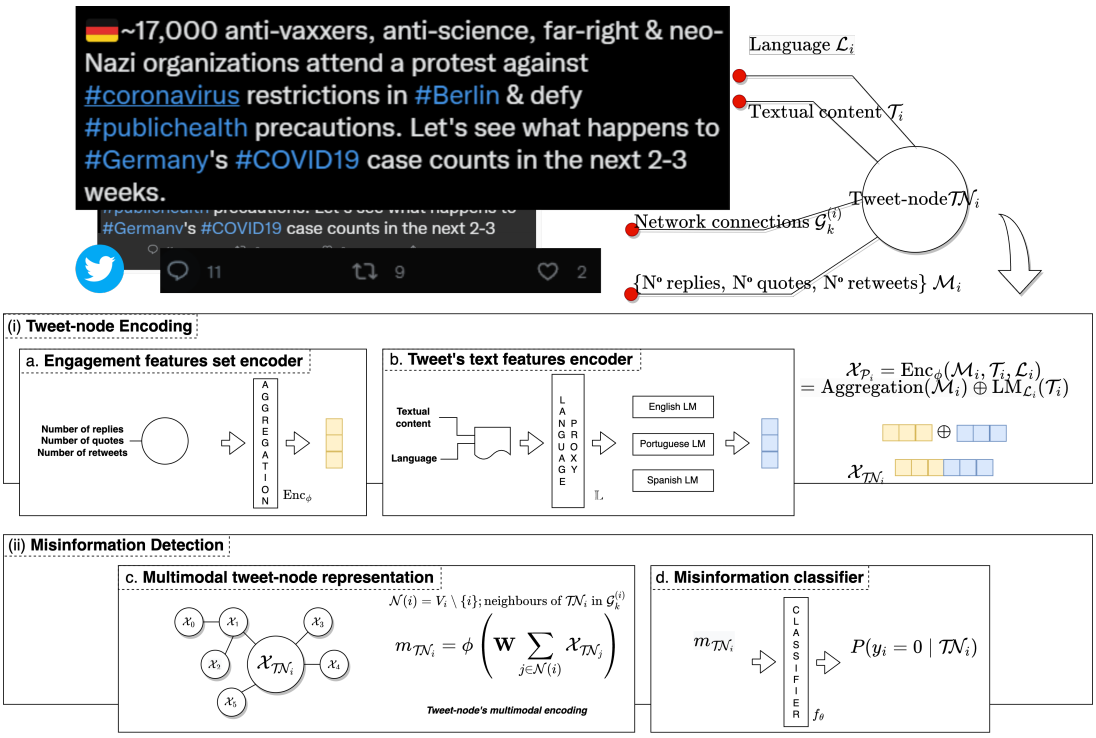


Figure 5: Architectural overview of the proposed mu2 framework instantiated for the Twitter/X ecosystem. Yellow and blue elements represent graph-based and text-based feature vectors, respectively.

a cross-entropy loss against y_i over $\mathcal{D}$ (Equation (3.6)).

3.3.2 mu2: Multimodal and Multilingual Framework for Misinformation Detection

The proposed framework consists of two main modules: *(i)* post-node encoding; *(ii)* misinformation detection. Moreover, module *(i)* is further partitioned into two components: a content encoder and a relational representation.

Figure 5 illustrates the instantiation of both modules – *post-node encoding* and *misinformation detection* – and the underlying relationship between them.

3.3.2.1 Post-node Encoding

This module is characterised by the transformation of a given post-node $\mathcal{P}_i$ into its corresponding vector representation.

Content Encoder Component Specifically, the component is divided into sub-components designed to address the heterogeneous nature of the data, as shown in Figure 5.i for meta-

data $\mathcal{M}_i$ and text $\mathcal{T}_i$. The engagement sub-component produces a vector representation of engagement features, including interaction counts for replies, quotes, and re-posts (Figure 5.i.a). The second part uses the language label $\mathcal{L}_i$ as a router, assigning the textual content $\mathcal{T}_i$ to the corresponding pre-trained language-specific model (LM) (Figure 5.i.b).

Finally, as described in Equation 3.2, both feature vectors are concatenated to form a unique multimodal feature representation $\mathcal{X}_{\mathcal{P}_i}$.

$$\mathcal{X}_{\mathcal{P}_i} = \text{Enc}_\phi(\mathcal{M}_i, \mathcal{T}_i, \mathcal{L}_i) = \text{Aggregation}(\mathcal{M}_i) \oplus \text{LM}_{\mathcal{L}_i}(\mathcal{T}_i). \tag{3.2}$$

Relational Representation Component This component enhances the post-node's multimodal feature representation $\mathcal{X}_{\mathcal{P}_i}$ with contextual information collected from the local network connections $\mathcal{G}_k^{(i)}$, as illustrated in Figure 5.ii.a. Specifically, these network connections are defined by the various relations a post maintains within a social network, including the posts that quote, reply to, or share it.

The combination of the multimodal feature representation with these network features produces a new vector representation termed the *multimodal post-node representation*. This approach allows the model to consider both the content of the post and its structural context within the social media environment. To achieve this, the learning process for the multimodal post-node representation $m_{\mathcal{P}_i}$ is defined according to Equation 3.3.

Given a set of post-node features $\mathbf{X} = \{\mathcal{X}_{\mathcal{P}_1}, \mathcal{X}_{\mathcal{P}_2}, \ldots, \mathcal{X}_{\mathcal{P}_N}\}$ with $\mathcal{X}_{\mathcal{P}_j} \in \mathbb{R}^{|\mathcal{X}_{\mathcal{P}_i}|}$, and letting $\mathcal{N}(i) = V_i \setminus \{i\}$ denote the set of neighbours of $\mathcal{P}_i$ in $\mathcal{G}_k^{(i)}$,

$$m_{\mathcal{P}_i} = \phi\left(\mathbf{W} \sum_{j \in \mathcal{N}(i)} \mathcal{X}_{\mathcal{P}_j}\right), \tag{3.3}$$

where $\phi(\cdot)$ is a non-linear activation function, and $\mathbf{W}$ is a linear transformation.

3.3.2.2 Misinformation Detection

Following the derivation of the multimodal post-node representation $m_{\mathcal{P}_i}$, the vector is passed through a classification layer to determine the veracity of the post. In this step, a softmax function is utilised to calculate the probability distribution based on the representation. Equation 3.4 defines this process, yielding the probability that the post $\mathcal{P}_i$ belongs to the misinformation class.

$$P(y_i = 0 \mid \mathcal{P}_i) \;=\; \big[\mathrm{softmax}(\mathbf{W}_{\mathrm{cls}}\, m_{\mathcal{P}_i})\big]_0 \;=\; \frac{\exp\big((\mathbf{W}_{\mathrm{cls}}\, m_{\mathcal{P}_i})_0\big)}{\sum_{c\in\{0,1\}} \exp\big((\mathbf{W}_{\mathrm{cls}}\, m_{\mathcal{P}_i})_c\big)}, \tag{3.4}$$

where $\mathbf{W}_{\mathrm{cls}}$ is the classification head that projects the post-node representation onto a 2-dimensional logit vector and $\big[\,\cdot\,\big]_0$ denotes the misinformation class selection. The predicted probability used in the loss is the complementary fact-class probability $\hat{y}_i = 1 - P(y_i = 0 \mid \mathcal{P}_i)$, consistent with the convention introduced in Section 3.3.1.

3.3.3 Framework Instantiation

This section details the mechanisms used to instantiate the previously introduced components illustrated in Figure 5. Here, we focus on Twitter posts, calling them *tweet-node*s.

Language Models We utilise pre-trained, language-specific models to encode the textual features of the tweets. Specifically, for English tweets we use the BERTweet[9] model (NGUYEN; VU; TUAN NGUYEN, 2020) as the encoder; Spanish tweets utilise RoBERTuito[10] (PÉREZ et al., 2022); and Portuguese tweets are encoded using BERTweet.BR[11] (CARNEIRO et al., 2025).

Graph-based Classifier For the misinformation detection task, we employ the Graph Attention Network (GAT) (VELIČKOVIĆ et al., 2018) architecture combined with a softmax classifier as defined in Equation 3.4.

Consequently, Equation 3.3 is computed by using the graph attention mechanism (VELIČKOVIĆ et al., 2018), which applies a masked self-attention mechanism where the graph topology acts as a structural mask restricting attention to each node's local neighbourhood $\mathcal{N}(i)$. The attention coefficient $\alpha_{ij} \in [0, 1]$ is the softmax-normalised attention score between $\mathcal{X}_{\mathcal{P}_i}$ and $\mathcal{X}_{\mathcal{P}_j}$, with raw scores produced by a learned function $e : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ where $d = |\mathcal{X}_{\mathcal{P}_i}|$.

$$m_{\mathcal{P}_i} = \phi\left(\mathbf{W} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}\, \mathcal{X}_{\mathcal{P}_j}\right). \tag{3.5}$$

Finally, to train the model, we utilise the negative log-likelihood loss function (binary cross-entropy) defined in Equation 3.6. The training objective aims to minimise the error between the true labels and the predicted outputs across the training set.

[9] https://huggingface.co/docs/transformers/model_doc/bertweet
[10] https://huggingface.co/pysentimiento/robertuito-base-cased
[11] https://huggingface.co/melll-uff/bertweetbr

Here, i corresponds to a sample from the training set of size N, y_i represents the true label, and $\hat{y}_i = f_\theta(\mathcal{P}_i)$ is the predicted probability for the sample.

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right), \tag{3.6}$$

3.4 Experimental Setup

We conduct a series of empirical evaluations to demonstrate the robustness and efficacy of mu2, our proposed multimodal and multilingual framework. In these evaluations, we specifically assess the framework's performance on the misinformation detection task.

We perform our evaluations on a benchmark that integrates multilingual and multi-topic texts as well as the network data required to contextualise them in the social media environment. At the time of this work, and to the best of our knowledge, the MuMiN dataset (NIELSEN; MCCONVILLE, 2022) was the only available resource that satisfied all the requirements of our evaluation framework. For the purposes of this study, we utilise the *MuMiN-small* variant, comprising 2,183 fact-checked claims and over 7 million associated tweets. We filtered the corpus to include the three most prevalent languages – English, Portuguese, and Spanish – incorporating four entity types (*Claim*, *Tweet*, *Reply*, and *User*) and six distinct relational edges (*Posted*, *Mentions*, *Retweeted*, *Quote_ Of*, *Reply_ To*, and *Discusses*).

The framework was implemented using PyTorch (PASZKE et al., 2019), and we provide the complete codebase[12] and implementation details to ensure reproducibility. Following the experimental protocol established by the dataset authors, our misinformation classifier employs GAT to capture relational information. Model optimisation was performed on a single Nvidia RTX4070 Mobile GPU utilising the Adam optimiser (KINGMA; BA, 2015) with a learning rate of 0.005 over 400 epochs and a 16-dimensional hidden layer. Textual features were projected from their native 768-dimensional embedding space to an 812-dimensional space via linear mapping to ensure dimensional alignment with the engagement feature representations.

The evaluations subsequently presented address the multilingual configuration whilst systematically comparing the impact of various modalities. Reported results are generated by 1000 bootstrap resampling with 95% confidence interval.

[12] https://github.com/vitornl/MeLLL-MuMiN-explainable

Table 1: Misinformation classification F1 for each language and modality.

GAT's input Feature representation	Multilingual	English	Spanish	Portuguese
Graph-based	0.9488 ± 0.0032	0.9419 ± 0.0045	0.9477 ± 0.0029	0.9920 ± 0.0038
Text-based	0.9487 ± 0.0032	0.9409 ± 0.0044	0.9486 ± 0.0028	0.9923 ± 0.0042
Multimodal	**0.9738 ± 0.0057**	**0.9738 ± 0.0053**	**0.9500 ± 0.0004**	**0.9951 ± 0.0053**

3.5 Results and Analysis

The empirical outcomes of the tweet classification task are detailed in Table 1, where the model performs binary classification into either the *misinformation* or *fact* labels. We evaluate three distinct feature sets – graph-based representations, textual embeddings, and a multimodal fusion of graph and textual representations – utilising the F1-score as our primary metric of predictive performance.

Across all evaluated languages, the multimodal feature vector achieves better overall performance in misinformation detection compared to isolated textual encodings, corroborating the findings reported by Nielsen *et al.* (NIELSEN; MCCONVILLE, 2022). Notably, our analysis reveals that textual features exhibit greater robustness than graph-based representations in scenarios of low-resource languages or cases involving sparse graphs, such as the Spanish and Portuguese subsets.

3.6 Conclusion and Discussions

In this chapter, we proposed mu2, a multimodal and multilingual framework for misinformation detection in social networks that integrates language-specific text encoders with graph attention networks over social network structure. Our experiments on the `MuMiN` dataset across English, Portuguese, and Spanish demonstrate that multimodal feature fusion consistently achieves the best classification performance, with a multilingual F1-score of 0.9738 ± 0.0057 and per language scores of 0.9738 ± 0.0053 (English), 0.9500 ± 0.0004 (Spanish), and 0.9951 ± 0.0053 (Portuguese). The particularly strong performance on the Portuguese subset (0.9951) suggests that the framework is especially effective when both textual and relational signals are well-represented. Notably, our cross-lingual analysis reveals that textual features exhibit greater robustness than graph-based representations in low-resource and sparse graph scenarios, whilst graph-based and text-based features perform comparably in isolation across all evaluated languages. This finding carries practical implications for deployment in multilingual settings, where network structure may

be incomplete or unavailable: text-based encoders can serve as a reliable baseline, with multimodal fusion providing consistent improvements when relational data is accessible. These findings directly address Research Question 1, showing that multimodal integration improves on the best single-modality baseline in every evaluated language, with the largest gain in the well-resourced English subset.

There are several promising directions for future work to build upon these findings. First, the framework's modular architecture naturally accommodates the integration of additional modalities, such as image and video content, which are increasingly prevalent in social media misinformation campaigns; we hypothesise that incorporating these richer data sources will improve detection performance, particularly in media-rich contexts. Second, the use of foundation models as feature extractors, replacing the current language-specific encoders, could improve the quality of the underlying representations and extend coverage to additional languages without requiring dedicated pre-trained models for each. Third, the temporal dynamics of social media content and network structures warrant investigation; adaptive approaches that account for evolving misinformation patterns could enhance the framework's applicability to real-world deployment scenarios. Fourth, a systematic evaluation of alternative multimodal aggregation strategies (such as attention-based fusion or gated mechanisms) would provide a more comprehensive understanding of how to optimally combine heterogeneous feature representations. Finally, cross-dataset transfer experiments would strengthen the generalisability of our findings beyond the `MuMiN` benchmark. The explainability dimension of mu2, which elucidates the features driving individual classification decisions, is addressed in Chapter 5.

3.7 Limitations

The findings reported in this chapter are subject to two categories of limitations, those inherited from the multimodal misinformation detection paradigm and those specific to the design and scope of mu2.

Integrating heterogeneous modalities (text, graph, metadata), *i.e.*, *feature aggregation*, into a unified representation remains an open methodological challenge. Simple strategies such as vector concatenation may not fully exploit the complementary information across modalities, whilst more advanced fusion mechanisms introduce additional complexity and computational cost. A second limitation is *temporal dynamics*. Social media content and network structures evolve rapidly, and static models trained on fixed snapshots may

not generalise to emerging misinformation patterns. Addressing this last point requires adaptive approaches that track shifts in both content and network topology, which the static evaluation protocol adopted here does not exercise.

Other limitations worth noting fall into two groups. The first concerns the *methodological scope of the evaluation*. Although mu2 supports several multimodal aggregation strategies in principle, our experiments evaluate only vector concatenation. Alternative strategies such as attention-based fusion or gated mechanisms could yield different performance characteristics and are deferred to future work, in part because exhaustively comparing fusion mechanisms is not the core contribution of this chapter and would dilute the analysis of the relational signal we set out to investigate. The second concerns *coverage*. The evaluation is conducted exclusively on the `MuMiN` dataset, restricted to three languages (English, Portuguese, and Spanish). `MuMiN` was chosen because it is the only widely adopted benchmark that combines multilingual claims with full social-network metadata of the kind mu2 requires. Equivalent multimodal multilingual benchmarks beyond `MuMiN` are not currently available, and the three language scope reflects the languages on which `MuMiN` provides sufficient annotated content. Validation on additional benchmarks with different annotation schemes, domains, and network structures, and on typologically diverse lower-resource languages, would strengthen the generalisability of the findings reported here.

4 Knowledge Graph-based Contrastive Reasoning with LLMs for Misinformation Detection

This chapter presents KG-CRAFT, a framework for automated fact-checking that leverages large language models to extract knowledge graphs from unstructured reports, formulate contrastive questions grounded in the extracted knowledge, and assess the veracity of a given claim. The natural language interpretation capabilities of KG-CRAFT are further explored in Chapter 6.

The main content is an extended version of the paper "KG-CRAFT: Knowledge Graph-based Contrastive Reasoning with LLMs for Enhancing Automated Fact-checking" published in the *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026* (LOURENÇO; PAES; WEYDE, et al., 2026).

4.1 Introduction and Background

This chapter addresses the claim verification task, a core component of the automated fact-checking pipeline (Section 2.3) that assesses the veracity of a given claim against available evidence (Definition 2.2). Whilst large language models (Section 2.6) have demonstrated notable capabilities across a wide range of natural language understanding tasks (AUGENSTEIN et al., 2024), their application to fact-checking remains limited to knowledge encoded during pre-training, which may be incomplete, outdated, or factually incorrect (WANG, Yuxia et al., 2024). At the same time, knowledge graph-based approaches (Section 2.5.2) to fact-checking offer structured and verifiable reasoning over entities and relations (QUDUS et al., 2025), yet they typically depend on pre-existing knowledge bases that may lack coverage for emerging claims. Therefore, the challenge lies in combining the reasoning capabilities of LLMs with the structured grounding provided by knowledge graphs in a manner that is both scalable and faithful to the available evidence.

We focus on an approach grounded in contrastive explanation (Section 2.4.4) (LIPTON, P., 1990), which seeks to understand why a particular fact holds *rather than* a plausible alternative, to propose a contrastive reasoning mechanism for claim verification. Recent work has demonstrated the utility of contrastive and counterfactual reasoning for model interpretability (JACOVI et al., 2021; PARANJAPE et al., 2021) and knowledge graph construction (CHEN; BERTOZZI, 2023; PAN, S. et al., 2024). However, these two lines of research have remained largely disconnected. A major limitation of existing LLM-based fact-checking methods is their reliance on either direct prompting or retrieved passages alone (CHEUNG; LAM, 2023; WANG, B. et al., 2024; XIE et al., 2025), without structured reasoning over the relationships between entities mentioned in the claim and the supporting evidence. We argue that grounding the reasoning process in knowledge graphs extracted from evidence reports, and formulating contrastive questions that probe alternative entity substitutions, can substantially improve the veracity assessment capabilities of pre-trained language models. This chapter addresses Research Question 2, which asks to what extent structured representations of evidence improve the veracity-assessment capabilities of large language models, compared with reasoning over unstructured evidence.

In this chapter, we propose `KG-CRAFT`, a framework that: *(i)* extracts knowledge graphs from unstructured evidence reports using LLMs; *(ii)* formulates and answers contrastive questions grounded in the extracted knowledge; and *(iii)* assesses claim veracity based on the distil contrastive evidence. We evaluate `KG-CRAFT` on two primary fact-checking benchmarks (`LIAR-RAW` (ALHINDI; PETRIDIS; MURESAN, 2018), six classes, and `RAWFC` (YANG, Z. et al., 2022), three classes) using multiple LLM backbones, and conduct additional cross-domain experiments on `SciFact` (WADDEN; LIN, et al., 2020) and `PubHealth` (KOTONYA; TONI, 2020a). Our experiments include comprehensive ablation studies examining the impact of KG-based versus LLM-generated contrastive questions, the effect of the number of questions, and the feasibility employing the framework into small language models. The main contributions of this chapter are as follows:

- We propose `KG-CRAFT`, a framework that integrates LLM-based knowledge graph extraction with contrastive reasoning for automated claim verification;
- We demonstrate that KG-grounded contrastive questions substantially outperform LLM-generated alternatives, with `KG-CRAFT` achieving state-of-the-art results on `LIAR-RAW` (73.87% F1) and `RAWFC` (81.58% F1); and
- We provide evidence that the structured reasoning produced by `KG-CRAFT` can be effectively distilled into small language models, with a 1.7B-parameter model

matching the performance of significantly larger naïve LLM baselines.

In the following sections, we first survey the related work on contrastive reasoning, knowledge graph construction, and automated fact-checking (Section 4.2). We then present the KG-CRAFT framework, detailing the knowledge graph extraction, contrastive reasoning, and claim verification components (Section 4.3). Subsequently, we describe the experimental setup (Section 4.4), report and analyse the main results and ablation studies (Section 4.5), present additional cross-domain experiments (Section 4.6), and conclude with a discussion of findings and future directions (Section 4.7).

4.2 Related Work

This section surveys the literature most relevant to the contributions of this chapter across contrastive reasoning, knowledge graph construction, and automated fact-checking. The discussion situates KG-CRAFT within the current methodological landscape and motivates the selection of the evaluation benchmarks adopted in section 4.4.

4.2.1 Contrastive Explanations and Reasoning

The principle of contrastive explanations is rooted in answering counterfactual why-questions by comparing an observed outcome with hypothetical alternatives (LIPTON, P., 1990; GUIDOTTI, 2024; VERMA, S. et al., 2024). Formally, contrastive explanations address "Why did P happen?" in terms of "Why did P happen rather than Q?", where P is an observed event and Q represents an alternative hypothesis (STEPIN et al., 2021). This framing establishes a minimum criterion whereby explanations must demonstrate why the observed event was more probable than its alternatives, thereby providing more discriminative information than non-contrastive accounts.

Recent work has applied these principles to improve model interpretability and reasoning quality. (JACOVI et al., 2021) projected inputs into a latent space that captures only features distinguishing potential decisions, enabling models to better identify which aspects support or contradict specific predictions. (PARANJAPE et al., 2021) extended this concept to commonsense reasoning tasks, demonstrating that pre-trained language models can generate contrastive explanations highlighting key differences between alternatives, improving both task performance and explanation faithfulness. These results evidence the effectiveness of contrastive reasoning across diverse tasks and its potential for settings

where distinguishing between competing interpretations is paramount.

In the context of fact-checking, assessing claim veracity inherently requires comparing what a claim asserts against what the available evidence supports. KG-CRAFT applies contrastive reasoning to this setting by leveraging knowledge graph structure to formulate contrastive questions that systematically compare the claim's assertions with alternative interpretations grounded in the evidence, thereby guiding the verification process through structured comparison rather than unconstrained generation.

4.2.2 Knowledge Graph Construction Using LLMs

A knowledge graph (KG) represents information as a graph structure, where nodes are *entities* connected by *relations*. Formally, $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathbb{C})$, where $\mathcal{T}$ contains triples (h, r, t) with entities $\{h, t\} \in \mathcal{E}$ and relations $r \in \mathcal{R}$.

Traditional approaches to knowledge graph construction have relied on manual curation by domain experts or rule-based information extraction pipelines (PAN, S. et al., 2024). Whilst these methods produce high-precision structured knowledge, they face inherent scalability limitations: manual curation is prohibitively labour-intensive for large corpora, and rule-based extractors generalise poorly across domains and linguistic variations. The shift towards neural information extraction methods alleviated some of these constraints, with supervised models trained on annotated corpora achieving improved entity and relation extraction performance; however, these approaches remain dependent on the availability of task-specific training data.

The emergence of LLMs has opened new avenues for automated KG construction, leveraging their broad linguistic knowledge to perform entity recognition and relation extraction in zero-shot or few-shot settings. (CHEN; BERTOZZI, 2023) demonstrated that LLMs can construct domain-specific knowledge graphs from unstructured text with minimal human intervention, whilst (ZHANG; SOH, 2024) proposed prompting strategies that guide LLMs to extract structured triples from diverse text sources. A comprehensive survey by (PAN, S. et al., 2024) systematically catalogues the landscape of LLM-augmented KG construction, and (ZHU et al., 2024) provide an empirical analysis showing that LLMs achieve competitive or superior performance in entity and relation extraction compared to specialised neural extractors.

KG-CRAFT leverages these capabilities by employing LLMs for KG extraction from claim-associated reports, using phased few-shot prompting to first identify entities, then

classify them into conceptual categories, and finally extract the relations that connect them (section 4.3.2). This approach enables the construction of claim-specific knowledge graphs without reliance on pre-existing knowledge bases, providing the structured foundation upon which the subsequent contrastive reasoning operates.

4.2.3 Automated Fact-Checking

Automated fact-checking (AFC) encompasses a range of computational methods designed to assess the veracity of claims against available evidence (BEKOULIS; PAPAGIANNOPOULOU; DELIGIANNIS, 2021). The methodologies surveyed in this subsection are organised into three strands: traditional encoder-based approaches that preceded the LLM era, knowledge graph-based methods that leverage structured representations for verification, and LLM-driven approaches that place large language models at the core of the fact-checking pipeline. This progression reflects a broader shift in the field from supervised classification with fixed representations towards more flexible, reasoning-oriented verification strategies.

4.2.3.1 Traditional Fact-Checking

Early approaches to AFC typically rely on text embeddings to represent claims and supporting documents, employing encoder-based architectures for classification. A foundational example is `dEFEND` (SHU; CUI, et al., 2019), which introduced a sentence-comment co-attention network that jointly analyses news content and user comments, demonstrating that social context can serve as a complementary signal for veracity assessment. Extending the thematic scope of AFC to public health, (KOTONYA; TONI, 2020a) developed `SBERT-FC` alongside the `PubHealth` dataset, providing both a domain-specific benchmark and an expanded explainability analysis. (ATANASOVA et al., 2020) proposed `GenFE` and `GenFE-MT`, which jointly optimise veracity prediction and explanation generation through multi-task learning, showing that the two objectives are mutually reinforcing.

Beyond single-document verification, `CofCED` (YANG, Z. et al., 2022) introduced a hierarchical architecture that combines an encoder with cascaded evidence selectors to process reports from multiple sources, enabling both coarse-to-fine claim verification and explanation generation. Whitehouse et al. (2022) integrate external knowledge bases into pre-trained models, demonstrating that incorporating Wikidata improves classification accuracy, particularly for political claims where the knowledge base provides timely and relevant contextual information.

4.2.3.2 Knowledge Graph-based Fact-Checking

Whilst the traditional approaches discussed above operate primarily over unstructured text, a complementary line of research explores the use of knowledge graphs as structured representations for claim verification. (CIAMPAGLIA et al., 2015) introduced one of the earliest computational approaches, demonstrating that the veracity of KG triple statements can be approximated by computing shortest paths between entity nodes in a knowledge graph derived from Wikipedia. (SHIRALKAR et al., 2017) extended this paradigm by framing fact-checking as a network flow problem, proposing the Knowledge Stream algorithm to identify evidential paths that support or refute a given triple. Their evaluation showed that the flow-based approach outperformed path-based methods across a range of real-world datasets.

These foundational approaches operate over pre-existing, large-scale knowledge graphs (*e.g.*, DBpedia, Wikidata) and evaluate claims expressed as structured triples. (KIM, J. et al., 2023) formalised this setting further by introducing `FactKG`, a benchmark of 108,000 natural language claims grounded in DBpedia, categorised across five reasoning types (one-hop, conjunction, existence, multi-hop, and negation) to enable systematic evaluation of KG-based verification methods. More recently, (CHEN, Y. et al., 2025) proposed `GraphCheck`, which extracts knowledge graph from text and incorporates them into an LLM-based verifier through instruction-style prompting, achieving strong results on scientific and health-domain benchmarks. A comprehensive survey of these and further knowledge graph-based fact-checking methodologies is provided by (QUDUS et al., 2025).

Whilst these works demonstrate the value of structured knowledge for claim verification, they predominantly rely on querying or traversing pre-constructed knowledge graphs to assess claims expressed as structured triples. In contrast, KG-CRAFT extracts knowledge graphs directly from unstructured reports associated with each claim and leverages the resulting structure to formulate contrastive questions that guide LLM-based reasoning, rather than performing direct graph traversal or embedding-based truth assessment.

4.2.3.3 Fact-checking Using LLMs

The knowledge graph-based approaches discussed in the preceding subsubsection increasingly incorporate LLMs as components – `GraphCheck` (CHEN, Y. et al., 2025), for instance, uses an LLM as its underlying verifier. This subsubsection shifts focus to approaches where the LLM serves as the *primary reasoning and verification engine*, rather than operating

alongside or in service of a structured knowledge representation.

The emergence of LLMs and their strong performance across diverse domains has led to a growing body of work that places them at the core of AFC. Whilst LLMs can provide answers to factual questions, their reliability is limited by the extent of their training data and their tendency to hallucinate factual statements (WANG, Yuxia et al., 2024; AUGENSTEIN et al., 2024). Several frameworks have been proposed to address these challenges of reliability and transparency.

Early work in this direction includes `FactLLaMA` (CHEUNG; LAM, 2023), which combined instruction-following with external evidence retrieval and demonstrated that augmenting LLMs with external knowledge sources improves fact-checking accuracy, particularly for claims related to recent events. `IKA` (GUO; ZENG, et al., 2023) builds a graph of positive and negative examples from labelled data to enhance both fact-checking and explanation capabilities. (LIU, H. et al., 2024) introduced `TELLER`, a dual-system framework that emphasises trustworthiness through the integration of human expertise and LLM capabilities. (WANG, B. et al., 2024) developed a defence-based framework that addresses the limitations of uncensored crowd wisdom by splitting evidence into competing narratives and leveraging LLMs for reasoned verification. More recently, `CorXFact` (TAN; ZOU; AW, 2025) proposes simulating human fact-checking principles by analysing claim-evidence correlations.

A popular strategy for improving LLM-based fact-checking is Retrieval-Augmented Generation (RAG). Examples include basic RAG-based architectures (SINGAL et al., 2024), RAG enhanced with fine-grained feedback mechanisms for optimising the retrieval task (ZHANG; GAO, 2024), and more advanced architectures focused on evidence retrieval and contrastive argument synthesis (YUE et al., 2024). Ferraz et al. (2026) further study RAG with small language models for fake news detection, showing that compact models can reach results competitive with established techniques while keeping the framework lightweight.

Other recent approaches address specific challenges within the fact-checking pipeline. `FIRE` (XIE et al., 2025) proposes an iterative approach aiming to improve the scalability and efficiency of claim verification, whilst (MENG et al., 2025) address the challenge of zero-day manipulated content through real-time contextual information retrieval.

KG-CRAFT differs from these LLM-centric approaches in a fundamental respect: rather than relying on the LLM's parametric knowledge or retrieved passages alone, it first constructs a structured knowledge graph from the evidence and then uses this graph

to formulate targeted contrastive questions that guide the LLM's reasoning. This design combines the structured verification strengths of knowledge graph-based methods with the flexible reasoning capabilities of LLMs, positioning KG-CRAFT at the intersection of the two paradigms reviewed in this and the preceding subsubsection.

4.2.3.4 Fact-Checking Datasets

This section reviews the principal datasets used to evaluate automated fact-checking methods, with particular emphasis on those adopted in the experimental evaluation of this chapter.

LIAR (WANG, 2017) One of the earliest large-scale benchmarks for fact-checking, `LIAR` comprises 12,836 short statements collected from PolitiFact and labelled with six fine-grained veracity classes (`pants-fire`, `false`, `barely-true`, `half-true`, `mostly-true`, `true`). The dataset provides metadata such as speaker identity, party affiliation, and context, but does not include the full reports used by fact-checkers to reach their rulings.

LIAR-PLUS (ALHINDI; PETRIDIS; MURESAN, 2018) Extending `LIAR`, `LIAR-PLUS` augments each claim with automatically extracted justification sentences from the corresponding PolitiFact ruling reports. These justifications serve as extractive explanations for the veracity label, enabling the development and evaluation of explainable fact-checking methods.

LIAR-RAW (YANG, Z. et al., 2022) Further extending `LIAR-PLUS`, `LIAR-RAW` provides the full-length ruling reports associated with each claim, rather than extracted justification sentences alone. This design supports methods that require access to complete report-level context for evidence distillation and verification.

RAWFC (YANG, Z. et al., 2022) Collected from Snopes, `RAWFC` comprises claims spanning diverse topics with three veracity classes (`false`, `half`, `true`) and their associated reports retrieved using claim keywords. Like `LIAR-RAW`, this dataset provides full reports, enabling evaluation of methods that operate over complete evidence documents.

PubHealth (KOTONYA; TONI, 2020a) Focusing on the health and public policy domain,

`PubHealth` contains claims sourced from fact-checking and news review websites, each associated with a veracity label and a journalist-written explanation. The dataset supports both veracity prediction and explanation generation tasks.

SciFact (WADDEN; LIN, et al., 2020) Targeting scientific claim verification, `SciFact` pairs scientific claims with abstracts from the research literature, annotated with supporting or refuting labels and rationale sentences. The dataset enables evaluation of verification methods on domain-specific claims requiring technical reasoning over structured evidence.

The experimental evaluation presented in this chapter adopts `LIAR-RAW` and `RAWFC` (section 4.4) as primary benchmarks, as their provision of full-length reports is essential to KG-CRAFT's knowledge graph extraction phase, which operates over complete evidence documents rather than pre-selected sentences. `PubHealth` and `SciFact` are additionally employed to assess cross-domain generalisation (section 4.6.2), evaluating KG-CRAFT's robustness when applied to shorter, domain-specific evidence.

4.3 Method

We introduce a novel approach that leverages knowledge graphs to ground contrastive reasoning and enhance LLM fact-checking capabilities: ***K**nowledge **G**raph-based **C**ontrastive **R**easoning for **A**utomatic **Fac**t **T** Verification* (KG-CRAFT). Our work focuses on the claim verification component of automated fact-checking, integrating contrastive reasoning into the verification process through the use of structured evidence to generate contextually relevant contrastive queries, thereby guiding more accurate claim classification. We begin with the claim verification task formulation.

4.3.1 Problem Statement

Let $\mathcal{C}$ be a claim with a set of associated *reports* $\mathcal{R}_\mathcal{C} = \{r_i\}_{i=1}^{|\mathcal{R}_\mathcal{C}|}$, where each report $r_i = (s_{i,1}, \ldots, s_{i,|r_i|})$ is a sequence of sentences (*i.e.*, a document). Optionally, sentence-level *evidence* annotations are given as a set of indices $\varepsilon_\mathcal{C} \subseteq \{(i,j)\}$; if unavailable, set $\varepsilon_\mathcal{C} = \varnothing$. The objective is to predict a veracity label $\mathcal{V}_\mathcal{C} \in \mathcal{Y}$, $|\mathcal{Y}| \geq 2$ via a verifier $f_\theta : (\mathcal{C}, \mathcal{R}_\mathcal{C}) \mapsto \mathcal{V}_\mathcal{C}$. Evidence $\varepsilon_\mathcal{C}$ (when present) is used for analysis but is not required by the formulation.

Next, we present the components of KG-CRAFT, depicted in Figure 6: knowledge graph construction from the textual input (the claim and its associated reports) (section 4.3.2);

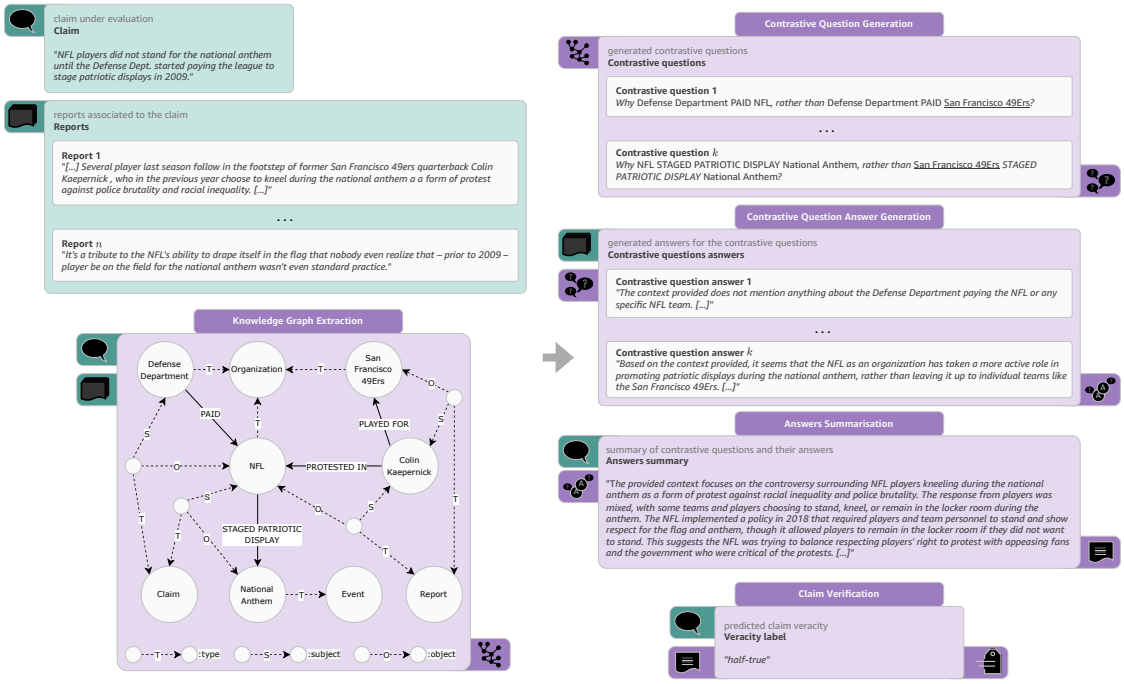


Figure 6: Overview of the KG-CRAFT framework for automated fact-checking. The method comprises three main phases: (1) knowledge graph extraction from the claim and associated reports, (2) contrastive reasoning, including contrastive question formulation, answer generation, and answer summarisation, and (3) claim veracity prediction.

contrastive reasoning, comprising contrastive question generation, answer generation, and prompt-based answer summarisation (section 4.3.3); and claim veracity verification (section 4.3.4).

4.3.2 Knowledge Graph Extraction

The first phase of KG-CRAFT extracts entities and relationships from $\mathcal{C}$ and $\mathcal{R}_\mathcal{C}$ and uses them to construct a knowledge representation of the input. We draw inspiration from prior work (ZHU et al., 2024; ZHANG; SOH, 2024) and leverage LLMs for knowledge graph construction. Through phased few-shot prompting, we instrument the LLM to first identify entities $\mathcal{E}$, then label them according to their conceptual categories $\mathbb{C}$, to give them semantic meaning and enable disambiguation. Next, we identify the relationships $\mathcal{R}$ that relate the identified entities (prompt details in appendix A.1.1), forming a set of triples $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. The unified resulting sets constitute the input knowledge graph $\mathcal{G}_{\mathcal{C},\mathcal{R}_\mathcal{C}} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathbb{C})$.

4.3.3 Contrastive Reasoning

The second phase of KG-CRAFT uses $\mathcal{G}_{\mathcal{C},\mathcal{R}_\mathcal{C}}$ to formulate and answer contrastive questions. It consists of: *(i)* formulating questions that contrast the claim's facts ($\mathcal{T}_{\text{claim}} \subseteq \mathcal{T}$) with the reports' facts ($\mathcal{T} - \mathcal{T}_{\text{claim}}$); *(ii)* answering the formulated questions using the reports ($\mathcal{R}_\mathcal{C}$); and *(iii)* summarising the question-answer pairs into a single self-contained paragraph.

Contrastive Question Formulation algorithm 1 describes the process to formulate contrastive questions. Given $\mathcal{G}_{\mathcal{C},\mathcal{R}_\mathcal{C}}$, the claim-specific triples $\mathcal{T}_{\text{claim}}$, and a maximum number K of desired questions, the algorithm creates the K most relevant and diverse contrastive questions. For each triple in the claim (consisting of *head* entity, relation, and *tail* entity), it first identifies the entity categories h_c and $h_t \in \mathbb{C}$ of the head and tail entities, respectively(algorithm 1, ll.6-8). Then, it creates two sets of contrastive entities: H_{contr}, containing alternative head entities of category h_c and T_{contr}, with alternative tail entities of category h_t (algorithm 1, ll.9-10). Using these sets, the algorithm generates questions by replacing either the original head or tail entity whilst maintaining the relation, following the pattern "Why [*original head*] rather than [alternative]?" or "Why [alternative] rather than [*original tail*]?" (algorithm 1, ll.11-14).

To ensure that the formulated questions are individually relevant and collectively diverse, we adopt a ranking strategy based on Maximal Marginal Relevance (MMR) (algorithm 1, l.20). For that, we first compute embeddings of the questions $\mathcal{Q}_{\text{Em}} = \{\textsc{Embedding}(q_i) \mid q_i \in \mathcal{Q}\}_{i=1}^{|\mathcal{Q}|}$, followed by the pairwise similarity matrix across all embeddings. The initial query embedding q_θ is selected as the embedding with the highest average similarity to all others, thereby ensuring a representative starting point.

We then iteratively construct the ranked set $\mathcal{Q}_{ranked}$ using the MMR procedure. At each iteration, the next embedding q_i is selected to maximise

$$\mathsf{q}_i = \underset{\mathsf{q} \in \mathcal{Q}_{\text{Em}} \setminus \mathcal{Q}_{ranked}}{\arg\max} \left[\text{Sim}(\mathsf{q}, \mathsf{q}_\theta) - \max_{\mathsf{q}' \in \mathcal{Q}_{ranked}} \text{Sim}(\mathsf{q}, \mathsf{q}') \right], \tag{4.1}$$

where $\text{Sim}(\cdot,\cdot)$ is the cosine distance, and $\mathcal{Q}_{ranked}$ is the set of already selected embeddings. The first term promotes relevance to the initial query q_θ, whilst the second penalises redundancy with respect to $\mathcal{Q}_{ranked}$. The process is repeated until all candidate questions are ranked into $\mathcal{Q}_{ranked}$. Finally, the algorithm returns the top K questions $\mathcal{Q}^{\mathsf{K}}_{ranked}$ (algorithm 1, ll.17-18).

Algorithm 1 Define Contrastive Questions

Require:

$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathbb{C})$ ▷ Claim and reports KG; entities $\mathcal{E}$, relations $\mathcal{R}$, triples $\mathcal{T}$, and entities' classes $\mathbb{C}$

$\mathcal{T}_{\text{claim}}$ ▷ Claim's extracted triples; $\mathcal{T}_{\text{claim}} \subseteq \mathcal{T}$

K ▷ Maximum # of Contrastive Questions

Ensure:

$\mathcal{Q}^{\mathsf{K}}_{ranked}$ ▷ Set of top K most relevant and diverse contrastive questions for a claim and its reports

$\mathcal{Q} \leftarrow \emptyset$

for all $\mathtt{t} = \{h,r,t\} \in \mathcal{T}_{\text{claim}}$ **do**

  $h_c \leftarrow \tau(h)$ ▷ where $\tau : \mathcal{E} \rightarrow \mathbb{C}$

  $t_c \leftarrow \tau(t)$

  $H_{contr} \leftarrow \{h \mid h \in \mathcal{E}, \tau(h) = h_c\} \setminus \{h\}$

  ▷ Set of entities of the same class as *head*

  $T_{contr} \leftarrow \{t \mid t \in \mathcal{E}, \tau(t) = t_c\} \setminus \{t\}$

  ▷ Set of entities of the same class as *tail*

  for all $h' \in H_{contr}$, $t' \in T_{contr}$ **do**

    $q_h \leftarrow \textsc{FormulateQuestion}(\mathtt{t}, h')$

    $q_t \leftarrow \textsc{FormulateQuestion}(\mathtt{t}, t')$

    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q_h, q_t\}$

  end for

end for

$\mathcal{Q}_{ranked} \leftarrow \textsc{ReRank}(\mathcal{Q})$ ▷ Rank contrastive questions based on eq. (4.1)

$\mathcal{Q}^{\mathsf{K}}_{ranked} \leftarrow (\mathcal{Q}_{ranked})_{1:K}$ ▷ Select first K elements from ranked set

return $\mathcal{Q}^{\mathsf{K}}_{ranked}$

Contrastive Question Answer Generation The second step of the process answers the previously formulated contrastive questions $\mathcal{Q}^{\mathsf{K}}_{ranked}$ through a structured query-response prompt mechanism p_{ag} (available in appendix A.1.2). An LLM analyses and generates information from the claim-associated reports – prompting it to answer independently the contrastive questions based on the set of reports $\mathcal{R}_{\mathcal{C}}$ associated with the claim. This process generates a set of answers $\tilde{\mathcal{A}} = \{LLM_{p_{\text{ag}}}(q, \mathcal{R}_{\mathcal{C}}) \mid q \in \mathcal{Q}^{\mathsf{K}}_{ranked}\}$, where each answer is directly derived from the reports, maintaining traceability between claim, reports, contrastive questions, and answers. Our goal by highlighting the contrastive elements in the generated answers is to highlight the key evidence supporting the claim's veracity during the reasoning process.

Answers Summarisation The final step of the Contrastive Reasoning process involves aggregating the question-answer pairs into a concise, evidence-based summary. From the claim $\mathcal{C}$ and the paired contrastive questions and answers from $\mathcal{Q}^{\mathsf{K}}_{ranked}$ and $\tilde{\mathcal{A}}$, respectively,

we prompt p_{as} (appendix A.1.3) to an LLM to generate a concise paragraph that relates all contrastive question-answer pairs. This summarisation step, represented as $A_{\mathcal{C}} = LLM_{p_{as}}(\mathcal{C}, \{(q_i, a_i) \mid q_i \in \mathcal{Q}^{\mathsf{K}}_{ranked}, a_i \in \tilde{\mathcal{A}}\})$, ensures that key contrasting elements and supporting evidence are presented in a structured summary, producing a distilled source of information. The resulting summary $A_{\mathcal{C}}$ emphasises key contrasting facts whilst abstracting non-essential information, preserving semantic links between critical evidence and the claim to create a focused source of verification.

4.3.4 Verification of Claim Veracity

The last phase of KG-CRAFT assesses the claim's veracity $\mathcal{V}_{\mathcal{C}}$. For that, we prompt p_{cv} (see appendix A.1.4) an LLM, containing the original claim $\mathcal{C}$ and the produced summary $A_{\mathcal{C}}$ as the only source of evidence, mitigating potential noise in the original reports $\mathcal{R}_{\mathcal{C}}$. The prompt p_{cv} also includes the possible labels and their descriptions. In this way, the LLM acts as a classifier to infer the claim's veracity, represented as $\mathcal{V}_{\mathcal{C}} = LLM_{p_{\text{cv}}}(\mathcal{C}, A_{\mathcal{C}})$, ensuring that the veracity assessment is based on the distilled evidence produced through our Contrastive Reasoning method.

4.4 Experimental Setup

Datasets We evaluated the proposed approach on two publicly available fact-checking datasets: `LIAR-RAW` and `RAWFC` (statistics are summarised in Table 2). `LIAR-RAW` (YANG, Z. et al., 2022) extends `LIAR-PLUS` (ALHINDI; PETRIDIS; MURESAN, 2018) and contains 12,590 fine-grained claims from PolitiFact[1] with six veracity classes (`pants-fire`, `false`, `barely-true`, `half-true`, `mostly-true`, `true`) along with their relevant reports. `RAWFC` (YANG, Z. et al., 2022) contains 2,012 claims collected from Snopes[2] on various topics with three veracity classes (`false`, `half`, `true`) and their associated reports retrieved using claim keywords. Both datasets split the claims into train, test, and validation sets containing, respectively, 80%, 10%, and 10% of the claims; `LIAR-RAW` and `RAWFC` are used in Sections 4.5.1 and 4.5.2.1. A modified two-class variant of `LIAR-RAW`, denoted `LIAR-RAW-binary` and used in Sections 4.5.2.2 and 4.5.2.3, is also reported in the table. This variant maps the original six labels into two classes – {`pants-fire`, `false`, `barely-true`} → `false` and {`half-true`, `mostly-true`, `true`} → `true` – whilst preserving the original train/test/validation split.

[1] www.politifact.com, [2] www.snopes.com

Table 2: Statistics of the LIAR-RAW and RAWFC datasets, including the LIAR-RAW-binary variant used in Sections 4.5.2.2 and 4.5.2.3. $|\mathcal{C}|_{ALL}$ denotes the total number of claims, and $|\mathcal{R}|_{avg}$ the average number of reports per claim.

Dataset	Statistics	Value
LIAR-RAW	pants-fire	1,013
	false	2,466
	barely-true	2,057
	half-true	2,594
	mostly-true	2,439
	true	2,021
	$\|\mathcal{C}\|_{ALL}$	12,590
	$\|\mathcal{R}\|_{avg}$	12.3
RAWFC	false	646
	half	671
	true	695
	$\|\mathcal{C}\|_{ALL}$	2,012
	$\|\mathcal{R}\|_{avg}$	21.0
LIAR-RAW-binary	false	5,536
	true	7,054

Comparisons We compare KG-CRAFT against other methods in three categories: *Traditional*, methods that do not use LLMs; *Naïve LLM*, direct application of LLMs without specialised prompts or reasoning strategies; and *Specialised LLM*, methods that use LLMs with significant adaptations, such as tuning or prompt engineering, or as part of complex architectures. Traditional approaches encompass dEFEND (SHU; CUI, et al., 2019), SBERT-FC (KOTONYA; TONI, 2020a), GenFE and GenFE-MT (ATANASOVA et al., 2020), and CofCED (YANG, Z. et al., 2022). Naïve LLM approaches encompass Llama 2 7B and ChatGPT 3.5 Turbo (WANG, B. et al., 2024), Claude 3.5 Sonnet, Claude 3.7 Sonnet, and Llama 3.3 70B prompted with the claim and related reports to verify the veracity of the claim. The Specialised LLM approaches encompasses FactLLAMA and FactLLAMA$_{\text{know}}$ (CHEUNG; LAM, 2023), and L-Defense$_{\text{LLAMA2}}$ and L-Defense$_{\text{ChatGPT}}$ (WANG, B. et al., 2024). FactLLAMA is a Low-Rank Adaptation (LoRA) (HU, E. J. et al., 2022) fine-tuned Llama 2 7B, and FactLLAMA$_{\text{know}}$ is the model augmented with external knowledge. L-Defense is a defence-based framework that leverages the wisdom of crowds to verify claim veracity using Llama 2 7B and ChatGPT 3.5 Turbo.

Implementation Details We extracted the KGs (Section 4.3.2) of both datasets utilising Claude 3 Haiku. Further, KG-CRAFT is instantiated and evaluated (Section 4.5.1) using Claude 3.5 Sonnet (KG-CRAFTC3.5), Claude 3.7 Sonnet (KG-CRAFTC3.7), and Llama 3.3 70B (KG-CRAFTL3.3).

Evaluation Metrics We adopted standard classification metrics: precision (Pr), recall (Re), and F1-score (F1). For all metrics, higher values indicate better performance.

4.5 Results and Analysis

This section presents the experimental evaluation of KG-CRAFT, organised around four research questions that progressively assess (i) whether the framework matches state-of-the-art on the headline benchmarks, (ii) whether the knowledge graph component is load-bearing relative to LLM-generated contrastive questions, (iii) whether the framework is robust to its main hyperparameter (the number of contrastive questions `K`), and (iv) whether the framework's reasoning can be effectively employed by smaller backbones.

The section is guided by the following research questions:

RQ1 How effective is KG-CRAFT in claim verification compared to state-of-the-art methods?

RQ2 How beneficial are KG-based contrastive questions compared to purely LLM-generated contrastive questions?

RQ3 What is the effect of the number of contrastive questions `K` on claim verification with KG-CRAFT?

RQ4 How effective is KG-CRAFT with Small Language Models (SLMs) compared to LLMs?

To answer these questions, we evaluate KG-CRAFT on two real-world fact-checking benchmarks and compare the results with baseline methods, and perform several ablation studies. Additional experiments and ablation studies can be found in section 4.6.

Table 3: Fact-checking results (%) on the RAWFC and LIAR-RAW datasets.

Method	LIAR-RAW			RAWFC		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**
Traditional						
dEFEND (SHU; CUI, et al., 2019)	23.09	18.56	17.51	44.93	43.26	44.07
SBERT-FC (KOTONYA; TONI, 2020a)	24.09	22.07	22.19	51.06	45.92	45.51
GenFE (ATANASOVA et al., 2020)	28.01	26.16	26.49	44.92	44.74	44.43
GenFE-MT (ATANASOVA et al., 2020)	18.55	19.90	15.15	45.64	45.27	45.08
CofCED (YANG, Z. et al., 2022)	29.48	29.55	28.93	52.99	50.99	51.07
EExpFND (WANG, J. et al., 2025)	29.91	29.93	29.58	54.43	53.49	53.61
Naïve LLM						
Llama 2 7B (WANG, B. et al., 2024)	17.11	17.37	15.14	37.30	38.03	36.77
ChatGPT 3.5 Turbo (WANG, B. et al., 2024)	25.41	27.33	25.11	47.72	48.62	44.43
Claude 3.5 Sonnet	29.25	27.83	28.52	54.85	55.04	54.94
Claude 3.7 Sonnet	28.26	26.63	27.42	55.90	57.12	56.50
Llama 3.3 70B	47.21	23.52	31.40	53.16	55.04	54.08
Specialised LLM						
FactLLAMA (CHEUNG; LAM, 2023)	32.32	31.57	29.98	53.76	54.00	53.76
$\text{FactLLAMA}_{\text{know}}$ (CHEUNG; LAM, 2023)	32.46	32.05	30.44	56.11	55.50	55.65
$\text{L-Defense}_{\text{LLAMA2}}$ (WANG, B. et al., 2024)	31.63	31.71	31.40	60.95	61.00	60.12
$\text{L-Defense}_{\text{ChatGPT}}$ (WANG, B. et al., 2024)	30.55	32.20	30.53	61.72	61.01	61.20
$\text{DelphiAgent}_{\text{gpt-4o-mini}}$ (XIONG et al., 2025)	32.79	22.33	26.03	68.53	63.95	64.68
$\text{DelphiAgent}_{\text{gpt-4o}}$ (XIONG et al., 2025)	31.33	28.36	28.36	68.05	68.03	68.04
KG-CRAFT (ours)						
KG-CRAFTC3.5	62.70	59.38	60.99	67.10	66.25	66.67
KG-CRAFTC3.7	**73.92**	**70.86**	**72.36**	**69.37**	**68.59**	**68.98**
KG-CRAFTL3.3	**77.38**	**70.67**	**73.87**	**81.63**	**81.53**	**81.58**

Note: The **best** and **second best** results are highlighted across each dataset and metric. KG-CRAFT results are significantly better than their Naïve LLM counterparts baseline with $p < 0.01$.

4.5.1 Claim Verification Evaluation

We first address RQ1 by evaluating the performance of KG-CRAFT in verifying claim veracity. As shown in Table 3, our method – in the instances KG-CRAFTC3.7 and KG-CRAFTL3.3 (all using five contrastive questions $\mathtt{K} = 5$) – consistently outperforms all other methods on both datasets. KG-CRAFTC3.5 outperforms all comparator methods except for DelphiAgent$_{\text{gpt-4o}}$) on the RAWFC dataset. KG-CRAFTL3.3 improves the F1-score by 44 percentage points (pp) on the LIAR-RAW dataset and 13 pp on the RAWFC dataset, compared to the second best performing methods, L-Defense (WANG, B. et al., 2024) and (DelphiAgent (XIONG et al., 2025)), respectively. Compared to their Naïve LLM counterparts, KG-CRAFTC3.5, KG-CRAFTC3.7, and KG-CRAFTL3.3, show F1 performance gains of 32, 44, and 42 pp on LIAR-RAW and 11, 12, and 27 pp on RAWFC.

4.5.2 Ablation Studies

To evaluate KG-CRAFT's components and reveal its strengths and limitations, we conduct three ablation studies. We first compare KG-based and LLM-generated contrastive questions to evaluate the benefits of structured question formulation (RQ2). Then, we analyse the effect of varying the number K of contrastive questions (RQ3). Finally, we assess whether KG-CRAFT boosts the performance of Small Language Models (SLMs) by comparing their results to Claude 3.7 Sonnet and KG-CRAFTC3.7 (RQ4).

4.5.2.1 Impact of Using LLM-generated Contrastive Questions

To answer RQ2, we replaced the Contrastive Question Formulation component of KG-CRAFT (section 4.3.3) by a few-shot prompt that, given the claim, reports, and examples, requests the LLM to generate $k = 5$ contrastive questions (prompt details in appendix A.1.5). Results depicted on table 4 show fact-checking F1-score, and macro and weighted AlignScore and RQUGE for the LIAR-RAW dataset.

AlignScore is a metric based on a general function of information alignment to perform automatic factual consistency evaluation of text pairs (ZHA et al., 2023). In our evaluation, we use AlignScore (here called macro AlignScore) to measure the information alignment between the text piece generated at the answer summarisation step (section 4.3.3) with the original claim. *RQUGE* is a reference-free evaluation metric for generated questions, based on the corresponding context and answer (MOHAMMADSHAHI et al., 2023). In our

evaluation, we use RQUGE (here called macro RQUGE) to measure the acceptability score of a formulated contrastive question (section 4.3.3), given a corresponding context (the claim) and an answer (the answer generated from the contrastive question section 4.3.3). We further extend AlignScore and RQUGE (here called weighted AlignScore and weighted RQUGE) to weight their results by the distance of the predicted claim veracity class from the ground-truth claim class. Based on the semantic distance between classes, we progressively assign each class an integer value ranging from 1 ($y_{\min}$) to $|\mathcal{Y}|$ ($y_{\max}$); for LIAR-RAW, this ranges from 1 (`pants-fire`) to 6 (`true`). We then define a per-claim distance weight that penalises predictions further from the gold class:

$$w_{\mathcal{C}} = 1 - \frac{(\mathcal{V}_{\mathcal{C}} - y)^2}{(y_{\max} - y_{\min})^2}, \tag{4.2}$$

where $\mathcal{V}_{\mathcal{C}}$ is the model's predicted class and y is the gold class. The weighted AlignScore is the product of $w_{\mathcal{C}}$ and the macro AlignScore between the answer summary $A_{\mathcal{C}}$ and the claim $\mathcal{C}$:

$$\text{AlignScore}_w = w_{\mathcal{C}} \times \text{AlignScore}(A_{\mathcal{C}}, \mathcal{C}). \tag{4.3}$$

Analogously, the weighted RQUGE re-weights the macro RQUGE over the set of contrastive questions:

$$\text{RQUGE}_w(\mathcal{C}) = w_{\mathcal{C}} \times \text{RQUGE}_{\text{macro}}(\mathcal{C}), \tag{4.4}$$

where $\text{RQUGE}_{\text{macro}}(\mathcal{C}) = \frac{1}{\mathsf{K}} \sum_{k=1}^{\mathsf{K}} \text{RQUGE}(q_k, \mathcal{C}, a_k)$ over the K formulated contrastive questions $q_k \in Q^{\mathsf{K}}_{ranked}$ and their generated answers $a_k \in \tilde{\mathcal{A}}$. Both weighted variants thereby penalise alignment and question-quality scores when the predicted veracity class deviates from the gold label, granting higher credit to predictions in semantically adjacent classes.

Table 4: Ablation study on the LIAR-RAW dataset comparing the quality of KG-based in contrast to LLM-based formulated contrastive questions. Metrics include Fact-checking F1-score, and macro (-M) and weighted (-W) values for AlignScore (AS) and RQUGE (RQ).

Metric	**LLM-based**		**KG-based**	
	C3.5	**L3.3**	**C3.5**	**L3.3**
FC F1	27.79	29.68	60.99	73.87
AS-M	30.96	29.51	41.71	40.32
AS-W	29.15	24.36	39.17	35.84
RQ-M	2.04	1.92	2.02	1.95
RQ-W	1.50	1.43	1.67	1.64

As depicted in table 4, the contrastive questions generated by the LLMs, results in lower information alignment (AlignScore) between the answer summary $A_{\mathcal{C}}$ and claim $\mathcal{C}$

and lower question acceptability score (RQUGE) between formulated contrastive questions Q^{K}_{ranked}, generated answer $\tilde{A}$, and claim $\mathcal{C}$. As a consequence, the lower information alignment and question acceptability score impact the fact-checking capability, resulting in a lower fact-checking F1-score.

4.5.2.2 Impact of the Number of Contrastive Questions

To address RQ3, we investigate how varying the number of contrastive questions k affects the performance of KG-CRAFT. We conduct this analysis on `LIAR-RAW-binary` (Section 4.4). We evaluate KG-CRAFTC3.5 using $k \in \{1,3,5,7,10\}$ contrastive questions, with $\mathsf{K} = 5$ serving as our baseline for relative performance comparison.

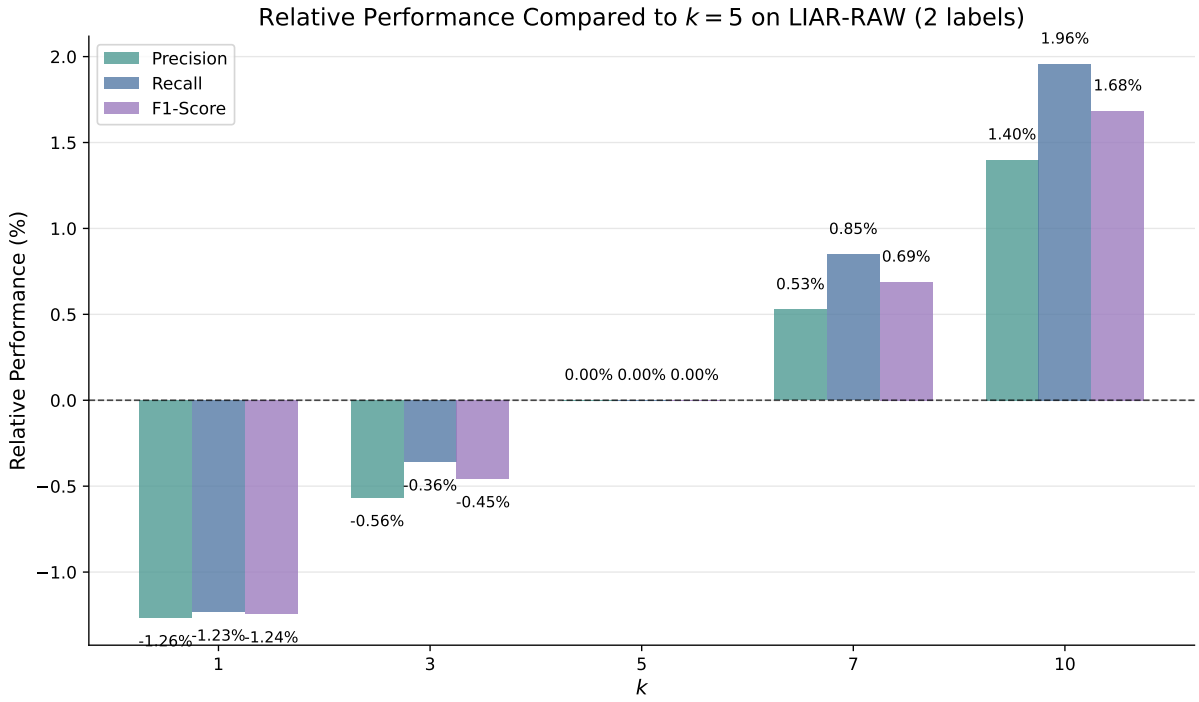


Figure 7: Impact of varying the number of contrastive questions k on fact-checking performance.

As depicted in Figure 7, increasing K generally improves performance metrics, but with progressively smaller gains. Specifically, compared to $\mathsf{K} = 5$, using fewer questions ($\mathsf{K} = 1$ or $\mathsf{K} = 3$) leads to decreased performance, with $\mathsf{K} = 1$ showing the most significant drops of 1.2 points across the metrics. Conversely, increasing the number of questions beyond $\mathsf{K} = 5$ yields small improvements, with $\mathsf{K} = 10$ showing the best relative gains of 1.6 points in F1. The relatively small performance differences (within ±2 points) indicate that our framework maintains robust performance across different K values, with $\mathsf{K} = 5$ representing an effective balance between performance and computational efficiency.

4.5.2.3 Using Small Language Models

We investigate whether our KG-based contrastive reasoning methodology can enhance the performance of Small Language Models (SLMs), as raised in RQ4. We evaluate four SLMs: two with less than 600M parameters (SmolLM2 135M (ALLAL et al., 2025) and Qwen3 0.6B (YANG, A. et al., 2025)) and two with less than 2B parameters (SmolLM2 1.7B (ALLAL et al., 2025) and Qwen3 1.7B (YANG, A. et al., 2025)). We compare their F1 performance against Naïve Claude 3.7 Sonnet and KG-CRAFTC3.7.

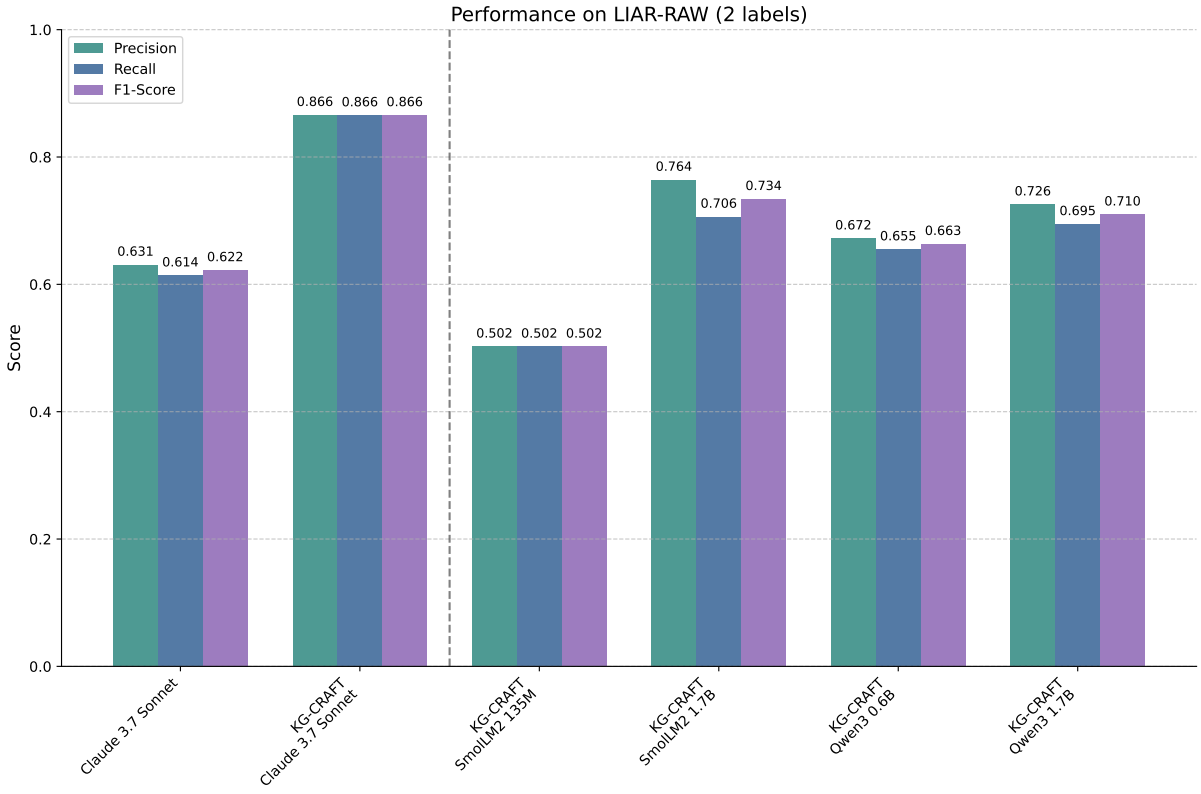


Figure 8: Performance comparison of Small Language Models incorporated within KG-CRAFT against larger models.

The results in fig. 8 show that our framework significantly enhances SLMs' fact-checking performance. Notably, SmolLM2 1.7B achieves an F1-Score of 73.40%, substantially outperforming Naïve Claude 3.7 Sonnet (62.22%). Even the smaller models show promising results, with Qwen3 0.6B reaching an F1-Score of 66.34%, which surpasses Naïve Claude 3.7 Sonnet despite its much smaller size.

These findings suggest that KG-CRAFT effectively compensates for the limitations of smaller models by supplying them with relevant verification cues. Applying KG-CRAFT significantly narrows the performance gap between SLMs and larger models, indicating that the explicit use of contrastive questions grounded in extracted KGs enhances smaller

Table 5: Comparative performance analysis (%) of KG-CRAFT and its key components on LIAR-RAW and RAWFC datasets.

Method	LIAR-RAW			RAWFC		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**
KG-CRAFT (ours)						
KG-CRAFTC3.5	62.70	59.38	60.99	67.10	66.25	66.67
KG-CRAFTC3.7	**73.92**	**70.86**	**72.36**	**69.37**	**68.59**	**68.98**
KG-CRAFTL3.3	**77.38**	**70.67**	**73.87**	**81.63**	**81.53**	**81.58**
Naïve LLM						
Claude 3.5 Sonnet	29.25	27.83	28.52	54.85	55.04	54.94
Claude 3.7 Sonnet	28.26	26.63	27.42	55.90	57.12	56.50
Llama 3.3 70B	47.21	23.52	31.40	53.16	55.04	54.08
LLM with KG augmentation (no contrastive reasoning)						
Claude 3.5 Sonnet	39.96	37.52	38.70	56.73	56.51	56.62
Claude 3.7 Sonnet	60.70	56.42	58.48	68.27	66.09	67.16
Llama 3.3 70B	61.50	38.11	47.06	60.64	58.54	59.57
LLM-based contrastive questions (no KG)						
Claude 3.5 Sonnet	34.54	23.25	27.79	65.40	66.56	65.97
Claude 3.7 Sonnet	30.20	27.05	28.54	67.65	68.57	68.11
Llama 3.3 70B	42.18	22.90	29.68	64.98	64.99	64.98

Note: The **best** and **second best** results are highlighted across each dataset and metric. *KG-CRAFT* and *Naïve LLM* results are same in table 3.

models' competitiveness.

4.6 Additional Experiments

4.6.1 Ablation Study

To evaluate the impact of KG-CRAFT's components and validate its effectiveness as a complete framework, we conducted a comprehensive ablation study on the LIAR-RAW and RAWFC datasets. Our analysis compares the full KG-CRAFT framework (proposed in Section 4.3 and results discussed in section 4.5.1) against three key architectural variations: Naïve LLM, an LLM augmented with only the knowledge graph (KG), and an LLM that generates contrastive questions without using the KG structure (presented and discussed in section 4.5.2.1). The results, presented in Table 5, provide a detailed breakdown of each component's contribution to the overall performance.

Naïve LLM The baselines Naïve LLM section (presented in Section 4.4 and results

discussed in section 4.5.1) shows the performance of aforementioned backbone LLMs when tasked with fact-checking using only the claim and its associated reports. This approach lacks any structured reasoning or pre-processing of the reports. The results for these models are notably lower than for KG-CRAFT, confirming the need for an enhanced reasoning mechanism.

LLM with KG augmentation (no contrastive reasoning) This ablation evaluates the backbone models augmented of the knowledge graphs extracted from the claim and reports, but bypassing the contrastive reasoning phase. The knowledge graph (KG) is provided as additional context to the LLM for veracity assessment. Comparing these results to the Naïve LLM baselines shows that augmenting the LLM with the extract (KG) does improve performance. However, the gains are marginal compared to the full KG-CRAFT framework, highlighting that the contrastive reasoning process is the primary driver of the significant performance increase. For instance, on the LIAR-RAW dataset, Llama 3.3 70B shows a 15pp increase in F1-score with KG augmentation, but the full KG-CRAFT framework provides a more substantial 42pp gain compared to the naïve baseline.

LLM-based Contrastive Questions (No KG) This ablation directly addresses RQ2 (Section 4.5.2.1) by replacing the KG-based question formulation with questions generated purely by the LLM using a few-shot prompt. The results indicate that this approach is less effective than our KG-based method. The F1-scores are significantly lower, for instance, with Llama 3.3 70B achieving only a 29.68% F1-score on LIAR-RAW, compared to the 73.87% F1-score of the full KG-CRAFT framework. This suggests that LLMs, when prompted to generate their own contrastive questions, often fail to create questions that are both contextually relevant and aligned with the provided reports, resulting in overall lower information alignment (AlignScore) between the answer summaries and the original claims and overall lower question quality (RQUGE) between formulated questions, generated answers, and the original claims (section 4.5.2.1). This finding reinforces the value of our knowledge graph approach for producing evidence-based questions.

KG-CRAFT (ours) The results of KG-CRAFT framework (proposed in Section 4.3 and results discussed in section 4.5.1), which combines all proposed components, consistently outperform all other variations, achieving a new state-of-the-art on both datasets (sec-

tion 4.5.1). This confirms that the combination of knowledge graph extraction with the proposed contrastive reasoning significantly enhances LLMs' fact-checking abilities.

4.6.2 Evaluation with Other Datasets

We further examine whether KG-CRAFT generalises to scenarios where claims are accompanied by fewer and shorter reports by evaluating on two benchmarks: **SciFact** (scientific abstracts) and **PubHealth** (health and policy claims).

4.6.2.1 Experimental Settings

Datasets SciFact dataset (WADDEN; LIN, et al., 2020) targets scientific claim verification from research paper abstracts. We used its validation set (also referred to as *dev* set), retaining claims with complete supporting or refuting evidence (label classes: `supports` and `refutes`). PubHealth (KOTONYA; TONI, 2020a) covers health-related and public policy claims. We use the test split, selecting claims that include the dataset's supporting context (label classes: `true`, `false`, `mixture`). In both settings, we strictly use the reports provided by each dataset, *i.e.*, no external retrieval (also known as gold evidence), so observed differences reflect reasoning rather than retrieval. Full results are in Tables table 6 and table 7.

Comparisons The performance of KG-CRAFT is benchmarked against seven competing methods, which include encoder-based classifiers, graph models, program-style pipelines, and LLM-based approaches. MLA (RoBERTa) (KRUENGKRAI; YAMAGISHI; WANG, 2021) proposes a sequence inference model which uses self-attention at both the token and sentence levels to capture information with a pre-LM encoder (RoBERTa). It also feeds static positional encodings into its multi-head attention mechanism. MULTIVERS (WADDEN; LO, et al., 2022) leverages a multi-task learning approach aiming at multi-evidence scientific verification and rationale extraction from abstracts. ProgramFC (PAN, L. et al., 2023) formulates fact verification as a programmatic pipeline with modular steps for evidence usage and decision making. PACAR (ZHAO et al., 2024) is a prompting-based approach that aggregates evidence with consistency-oriented reasoning. GraphFC (HUANG, Y. et al., 2025) encodes the report structure with an explicit graph and graph reasoning components. CO-GAT (ELECTRA) (LAN et al., 2025) applies graph attention over scientific evidence with a pre-LLM encoder (ELECTRA). GraphCheck (CHEN, Y. et al., 2025) is a recent LLM-

Table 6: Performance comparison on the SciFact dataset (%).

Method	Pr	Re	F1
Previous Methods			
MLA (RoBERTa) (KRUENGKRAI; YAMAGISHI; WANG, 2021)[a]	**80.62**	49.76	61.54
MULTIVERS (WADDEN; LO, et al., 2022)	73.80	71.20	72.50
ProgramFC (PAN, L. et al., 2023)[b]	-	-	71.82
PACAR (ZHAO et al., 2024)	-	-	75.06
GraphFC (HUANG, Y. et al., 2025)	-	-	**87.37**
CO-GAT (ELECTRA) (LAN et al., 2025)[d]	79.58	54.07	64.39
KG-CRAFT (ours)			
KG-CRAFT$_{C3.5}$	76.50	**75.18**	75.83
KG-CRAFT$_{L3.3}$	**84.58**	**81.53**	**83.03**

Note: The **best** and **second best** results are highlighted across each dataset and metric. All reported results use the evidence provided by the dataset; thus, no external source is referenced. [a] Results taken from (LAN et al., 2025). [b] Results taken from (ZHAO et al., 2024). [c] Results taken from (HUANG, Y. et al., 2025). [d] CO-GAT (ELECTRA) reported results use the large model and abstract-level evidence settings. MULTIVERS reported results use the full and abstract-level evidence settings.

Table 7: Performance comparison on the PubHealth and SciFact datasets (%).

Method	PubHealth				SciFact			
	BAcc	Pr	Re	F1	BAcc	Pr	Re	F1
GraphCheck$_{L3.3}$ (CHEN, Y. et al., 2025)	73.60	-	-	-	**89.40**	-	-	-
GraphCheck$_{Qwen\ 72B}$ (CHEN, Y. et al., 2025)	71.70	-	-	-	86.40	-	-	-
KG-CRAFT (ours)								
KG-CRAFTC3.5	61.02	76.50	75.18	75.83	75.17	76.50	75.18	75.83
KG-CRAFTL3.3	**78.66**	72.82	73.89	73.35	81.52	84.58	81.53	83.03

Note: The **best** results for the Balanced Accuracy (BAcc) are highlighted across each dataset. KG-CRAFT Precision, Recall, and F1-score are reported for completeness; they are not being compared with GraphCheck results.

based verifier that incorporates lightweight graph signals and instruction-style prompting. We report KG-CRAFT with two backbones: KG-CRAFT$_{C3.5}$ and KG-CRAFT$_{L3.3}$, with the same settings as reported in our main experiments (Section 4.4).

Evaluation Metrics As in our main experiments (Section 4.4), we report precision (Pr), recall (Re), and F1-score (F1) results. In addition, to compare with GraphCheck, we report balanced accuracy (BAcc). For all metrics, higher values indicate better performance.

4.6.2.2 Results

On SciFact, KG-CRAFT$_{\text{L3.3}}$ obtains 83.03 F1 (Pr/Re: 84.58/81.53), outperforming five out of the six competing techniques (*e.g.*, MULTIVERS 72.50 and ProgramFC 71.82; more than 10pp difference) whilst presenting competing results with the best performing method (GraphFC 87.37; less than 5pp) (Table 6). Also, in SciFact, GraphCheck$_{\text{L3.3}}$ obtained 89.40 BAcc, outperforming KG-CRAFT$_{\text{L3.3}}$ in 7.88pp (Table 7). On PubHealth, KG-CRAFT$_{\text{L3.3}}$ achieves the best BAcc at 78.66, surpassing GraphCheck$_{\text{L3.3}}$ (73.60; 5.06 pp) and GraphCheck$_{\text{Qwen 72B}}$ (71.70; 6.96 pp) (Table 7). Whilst KG-CRAFT$_{\text{C3.5}}$ is consistently weaker than KG-CRAFT$_{\text{L3.3}}$ results, where reported for completeness.

In both datasets, reports associated with each claim are shorter, we observe less diverse extracted KGs, *i.e.*, fewer entities and relations, which directly constrains the space of type-consistent substitutions and thus the capacity to formulate diverse contrastive questions. This likely contributes to the residual gap to GraphFC on SciFact, despite strong overall results and cross-domain robustness evidenced by PubHealth.

4.7 Conclusion and Discussions

In this chapter, we proposed `KG-CRAFT`, a framework for automated claim verification that integrates LLM-based knowledge graph extraction with contrastive reasoning. Our experiments on two primary benchmarks demonstrate that `KG-CRAFT` achieves state-of-the-art performance, with F1-scores of 73.87% on `LIAR-RAW` and 81.58% on `RAWFC`, representing improvements of +44 and +13 percentage points over the best prior methods, respectively. Our ablation studies reveal that KG-grounded contrastive questions are essential to this performance: replacing them with LLM-generated alternatives results in a dramatic decline (from 73.87% to 29.68% F1 on `LIAR-RAW`), confirming that structured knowledge grounding is the primary driver of the framework's effectiveness. These findings directly address Research Question 2, showing that structured representations of evidence substantially improve LLM veracity assessment over reasoning that operates on unstructured text alone. The framework exhibits robust performance across the number of contrastive questions (K), with differences within ± 2 percentage points for $K \in \{1, 3, 5, 7, 10\}$, and $K = 5$ providing an effective balance between performance and computational cost. Furthermore, cross-domain experiments on `SciFact` (F1 83.03) and `PubHealth` (BAcc 78.66) demonstrate the generalisability of the approach beyond political fact-checking. Notably, the structured reasoning produced by `KG-CRAFT` can be effectively distilled into small language models:

SmolLM2 1.7B achieves 73.40% F1, substantially outperforming naïve Claude 3.7 Sonnet (62.22%) despite being orders of magnitude smaller, suggesting that the framework's contrastive evidence summaries serve as an effective form of knowledge distillation.

There are several promising directions for future work to build upon these findings. First, the contrastive answer summaries produced by `KG-CRAFT` provide a natural foundation for generating human-readable explanations of verification decisions; exploring this direction would address the broader goal of explainable automated fact-checking, as further investigated in Chapter 6. Second, extending the evaluation to additional domains beyond politics, health, and science would provide a more comprehensive understanding of the framework's cross-domain capabilities, particularly in domains where evidence reports exhibit different structural characteristics. Third, the current pipeline relies on a fixed set of LLMs for intermediate tasks (knowledge graph extraction, answer generation, summarisation); investigating the impact of alternative or fine-tuned models at each stage could yield further improvements. Fourth, a deeper qualitative analysis of the intermediate components, particularly the quality of the extracted knowledge graphs and the relevance of the generated contrastive questions, would provide valuable insights into the framework's failure modes and inform targeted improvements. Fifth, the demonstrated effectiveness of SLM distillation motivates further investigation into lightweight deployment configurations, including quantised models and on-device inference, to broaden accessibility. Finally, whilst the framework is designed to be language-agnostic, extending the evaluation to non-English languages would validate this claim and identify potential limitations in language-specific intermediate components such as entity linking and relation extraction.

4.8 Limitations

The findings reported in this chapter are subject to two categories of limitations, those inherited from the LLM-based fact-checking paradigm and those specific to the design and evaluation of KG-CRAFT.

The first general limitation concerns *hallucination and factual reliability.* Large language models are known to produce fluent yet factually incorrect outputs, and KG-CRAFT chains LLM calls across knowledge graph extraction, contrastive question answering, and veracity verification, which can lead to an error introduced at any stage can propagate downstream and mislead the final assessment. The risk is particularly severe in automated fact-checking, where system outputs may be interpreted as authoritative assessments of

veracity. *Evidence oversimplification* is a related limitation, which may affect the answer summarisation step described in Section 4.3.3. When aggregating the contrastive question-answer pairs into a concise summary, or during any process that involves decisions about what to retain and what to discard, if the summarisation prompt is miscalibrated, critical contextual information can be lost, biasing the downstream verification. *Computational cost* and *reproducibility under model evolution* are two further limitations that affect the practical applicability of LLM-based pipelines. Specifically associated to KG-CRAFT multiple chained LLM calls per claim, limits practitioners without access to large-scale inference infrastructure, although the distillation evidence and its usage in SLMs reported in Section 4.5.2.3 offers a partial mitigation. Moreover, since the rapid succession of LLM versions means that absolute numbers reported against a particular backbone (*e.g.*, Claude 3.5 Sonnet) may not reproduce against a future iteration of the same model family. A further risk specific to LLM-based evaluation is *benchmark contamination*, where the backbones may have encountered these datasets during pre-training. Whilst this possibility cannot be entirely excluded, the naïve-LLM baselines (Section 4.5.1) run the same backbones without our method, so the consistent gains of KG-CRAFT over them are attributable to the proposed reasoning mechanism rather than to memorised labels.

Four limitations are specific to this chapter. First, we report quantitative end-to-end performance but do not *qualitatively verify the intermediate outputs* produced by the pipeline (*e.g.*, the extracted knowledge graphs, the formulated contrastive questions, and the answer summaries). These intermediate artefacts are central to the framework's reasoning chain, and a systematic quality evaluation of each component output (*e.g.*, sampling extracted KGs and rating their precision and completeness, or sampling contrastive questions and rating their consistency and relevance) would substantially strengthen confidence in the framework's behaviour beyond what end-to-end F1 reveals. Second, the experiments rely on a *fixed LLM configuration*. The use of Claude 3 Haiku for KG extraction, Claude 3.5/3.7 Sonnet and Llama 3.3 70B for verification, and SmolLM2 / Qwen3 variants for the SLM analysis, were chosen for accessibility and cost-performance balance rather than via systematic search. Thus, alternative models, larger backbones, or fine-tuned variants could yield meaningfully different results. Prior work uses different LLM families for the LLM-based comparators, thus the gap between methods may partly reflect backbone capability rather than methodological substance. However, the reported results in Section 4.5.1 include the naïve LLM backbones, allowing for understanding the contributions proposed by KG-CRAFT. Third, the primary evaluation is conducted on two English-language datasets (`LIAR-RAW` and `RAWFC`). Whilst the `SciFact` and `PubHealth`

cross-domain experiments (Section 4.6.2) demonstrate that the framework generalises beyond political fact-checking, validation on additional languages remains an open direction, and intermediate components such as entity linking and relation extraction may exhibit language-specific failure modes. Fourth, when evidence reports are short (as in `SciFact`), the extracted KGs are sparser and the type-consistent substitution space is correspondingly smaller, constraining the diversity of formulable contrastive questions, exposing a *report-length sensitivity*. The framework's effectiveness is therefore partially dependent on the richness of the available evidence, with implications for any future deployment over short-form claims (*e.g.*, social-media posts) where evidence is constrained at retrieval time.

Part II

Explaining Misinformation Detection

5 Explainable Text and Graph Learning for Misinformation Detection in Social Networks

This chapter presents mu2X, an extension of the mu2 framework introduced in Chapter 3, augmented with a Classification Explainability Module that provides post-hoc, feature-level explanations for individual classification decisions.

The main content is an extended version of the paper "*Exploring Content and Social Connections of Fake News with Explainable Text and Graph Learning*" (LOURENÇO; PAES; WEYDE, 2025) published in the *35th Brazilian Conference on Intelligent Systems, BRACIS 2025*.

5.1 Introduction and Background

Whilst automated misinformation (Definition 2.2) detection systems have demonstrated increasingly strong predictive performance (Chapter 3), their adoption in high-stakes decision-making contexts remains limited by a fundamental challenge: the opacity of the underlying models. As discussed in Section 2.4, Explainable Artificial Intelligence (Definition 2.6) seeks to address this limitation by providing mechanisms that render model decisions transparent and understandable to diverse stakeholders. In the context of misinformation detection, the need for explainability is particularly important; content moderators, fact-checkers, and end users require not only accurate predictions but also a comprehensible account of which features led a system to flag a given piece of content as misinformation (SHU; CUI, et al., 2019; KOTONYA; TONI, 2020a). Without such explanations, automated systems risk undermining the very trust they are designed to protect (GILPIN et al., 2018).

We focus on feature-level explainability for multimodal misinformation detection, aiming to identify which specific features (textual, relational, or metadata) drive individual

classification decisions, thereby providing interpretation (Definition 2.7) on which features contributes most for the system's classification. Existing approaches to explainable misinformation detection have predominantly relied on content-based or user-comment-based explanations (SHU; CUI, et al., 2019; CUI; SHU, et al., 2019; KOU; ZHANG, et al., 2020; KOU; SHANG, et al., 2022). A major limitation of these methods, however, is that they explain the decision in terms of the content a user sees, rather than the specific model features (textual, relational, and metadata) that drive the classifier's prediction. Moreover, the evaluation of explanation quality in this domain has largely been confined to interpretability assessments, with limited attention to the trustworthiness and robustness of the generated explanations. We argue that a comprehensive evaluation of explainability must address all three dimensions, ensuring that explanations are not only interpretable but also faithful to the model's decision process and resilient to perturbations.

This chapter addresses Research Question 3, which asks how the integration of multimodal features affects feature-level explanation quality across interpretability, trustworthiness, and robustness. We investigate this empirically by comparing graph-based, text-based, and multimodal feature representations across all three dimensions. We propose mu2X, an extension of the mu2 framework (Chapter 3) that augments the classification pipeline with a Classification Explainability Module employing GraphLIME (HUANG, Q. et al., 2023) for graph-based features and Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) for text-based features. We evaluate mu2X on the `MuMiN` dataset (NIELSEN; MCCONVILLE, 2022) across three languages (English, Portuguese, and Spanish), and propose two novel evaluation protocols to assess explanation trustworthiness and robustness alongside standard interpretability analysis. The main contributions of this chapter are as follows:

- We propose mu2X, a multimodal and multilingual explainable framework that extends mu2 with post-hoc, feature-level explanations for misinformation detection, combining GraphLIME for relational features with Integrated Gradients for textual features;
- We introduce two novel evaluation protocols for assessing the trustworthiness and robustness of feature-level explanations, complementing standard interpretability analysis; and
- We provide empirical evidence that multimodal explanations yield the most stable and trustworthy results, whilst text-based explanations exhibit the greatest robustness to noisy feature injection across all evaluated languages.

In the following sections, we first survey the related work on explainable approaches to misinformation detection (Section 5.2). We then present the mu2X framework, detailing the problem statement, the Classification Explainability Module, and the framework instantiation (Section 5.3). Subsequently, we describe the experimental setup, including the two proposed evaluation protocols (Section 5.4), report and analyse the results across interpretability, trustworthiness, and robustness (Section 5.5), and conclude with a discussion of findings and future directions (Section 5.6).

5.2 Related Work

This section surveys the literature most relevant to the contributions of this chapter, focusing on explainable approaches to multimodal misinformation detection. The discussion situates mu2X within literature and identifies the key limitations of prior work that motivate the explainability framework proposed in the following sections.

5.2.1 Explainable Multimodal Misinformation Detection

The process involved in assessing the veracity of social media posts or news articles extends beyond simple classification. Rather, the provision of interpretable explanations is of paramount importance to foster human trust in automated systems (MINH et al., 2022). To this end, `dEFEND` (CUI; SHU, et al., 2019; SHU; CUI, et al., 2019) pioneered as one of the first explainable frameworks for fake news detection. The primary objective of this methodology is the categorisation of news articles and associated comments into veracity labels. The authors proposed a multi-layered hierarchical encoding mechanism that synthesises news content with user commentary to perform the classification task. Whilst restricted to the textual modality, this work established a foundation by ranking news sentences and comments based on check-worthiness, thereby offering a sentence-level justification for the final classification. Following the development of `dEFEND`, (KOU; ZHANG, et al., 2020) proposed `ExFaux`, an approach specialised for explainable fauxtography detection. This approach differs from previous work by using a multimodal architecture that includes both images and text, whilst the explanations are provided by identifying specific parts of the content that led to the decision. `DGExplain` (SHANG et al., 2022) is an explainable multimodal framework designed for COVID-19 misinformation detection using text and images. In addition to its focus on COVID-19, this work extends the `ExFaux` architecture by incorporating user comments as a distinct feature set. The explanations generated

by this framework can be divided into comment-based and content-based categories: the former leverage social discourse to support the veracity decision, whilst the latter identify specific textual or visual elements as the primary drivers of the classification.

We argue that whilst the utilisation of users comments and content as explanatory tools eases the correlation of elements to improve interpretability, such approaches often do not identify which of the model's own features drove the classification. Therefore, this remains a significant research gap in attributing the decision to the specific features that drive the classification outcome. In this chapter, we aim to leverage the interpretability of these explanations, whilst also highlighting the specific features that drive the classification behaviour. Such feature attribution is commonly performed with model-agnostic methods such as LIME (RIBEIRO; SINGH; GUESTRIN, 2016) and SHAP (LUNDBERG; LEE, 2017), which estimate the contribution of individual input features to a prediction. mu2X adopts this post-hoc, feature-level paradigm, instantiated with GraphLIME (HUANG, Q. et al., 2023) and Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017), as detailed later in this chapter in Section 5.3.

`HC-COVID` Kou, Shang, et al. (2022) first integrates external knowledge bases to provide explanations for misinformation detection within the context of COVID-19 social media claims. Their work utilises a crowdsourced knowledge graph containing both general and domain-specific trustworthy information related to the pandemic. Moreover, the authors proposed a novel hierarchical attention mechanism designed to leverage this structured knowledge, *i.e.*, the knowledge graph, to detect misinformation in social media posts. In contrast to previous approaches that rely on parts of the post content or user comments as explanations, `HC-COVID` utilises previously accredited external knowledge to infer the veracity and provide a grounded justification for its classification decisions.

The use of previously audited information significantly enhances the interpretability of automated decision-making processes for the end user, thereby strengthening perceived reliability. However, similar to the aforementioned approaches, user-centric interpretability faces challenges when viewed from different perspectives, such as regulatory bodies that require an understanding of the features or mechanisms that led to the automated decision. Another limitation of this work is the scalability of the proposed solution to topic-agnostic misinformation detection. Whilst the crowdsourced knowledge graph provides expressibility for domain-specific entities (*e.g.*, events, pharmaceuticals, symptoms, among others) alongside their semantic relations and domain rules, the manual formulation of these graphs for wide variety of topics remains largely impractical. The many human behaviours

that lead topics to emerge in social networks, including their endurance and propagation patterns, requires a more flexible approach, as the current reliance on static, domain-specific graphs fundamentally limits the feasible scope of the solution.

Kotonya and Toni (2020a) tailor a domain-specific framework for the automated fact-checking of public health claims, utilising a hybrid of extractive, *i.e.*, explanations for truthfulness predictions corresponding to the inputs provided to the system, and abstractive summarisation techniques to generate explanations. The method identifies salient evidence by isolating sentences that support the veracity prediction, employing pre-trained language models to perform the extraction phase. As a second step, they employ an abstractive summarisation technique to create a human-readable explanation that combines these sentences into a coherent summary. This dual approach facilitates the generation of explanations accessible to a general audience whilst maintaining the facts of the underlying evidence. Furthermore, the authors performed one of the first human-centric evaluations to assess the quality of the generated explanations.

We highlight this contribution as one of the first to emphasise the interpretability of automated explanations. The work innovates from a method perspective by generating natural language explanations, whilst also introducing an evaluation methodology where human subjects annotate these outputs, *i.e.*, generated explanations, based on predefined qualitative criteria. However, one of the main limitations lies in the scale of the human evaluation, which utilised a restricted sample of 25 instances and five annotators tasked with answering two questions within a closed set of responses. We argue that such a constrained evaluation framework is susceptible to human biases, including confirmation and label biases, which may compromise the generalisability of the findings.

Finally, Yao et al. (2023) introduced a multimodal, multi-topical dataset of articles, along with an end-to-end framework designed to automate misinformation detection and provide interpretability. Their explainable approach comprises three primary stages: a multimodal evidence retrieval module, a claim verification module, and an explanation generation phase. During the initial stage, relevant evidence to either corroborate or refute the claim is retrieved from both textual and visual content. Subsequently, the system employs pre-trained language models to analyse the claim in conjunction with the retrieved evidence to determine its veracity, categorising it as supported, refuted, or containing insufficient information, thus determining the truthfulness of the claim. To ensure the trustworthiness of the prediction, the method generates ruling statements by leveraging the input claims, the predicted veracity label, and supporting textual evidence.

Closest to our work, Yao et al. (2023) integrates multiple relevant aspects for automated misinformation detection, including leveraging multimodal features of articles and generating trustworthy explanations. Despite these contributions, the proposed framework and associated dataset are restricted to the English language. Additionally, the framework lacks a specific focus on regarding and evaluating the interpretability of the provided explanations. The breadth of these explainable misinformation detection methods, spanning content-, comment-, and knowledge-based explanations, is surveyed by Gongane, Munot, and Anuse (2024), who catalogue explainable AI techniques for fake news and hate speech detection on social media.

Beyond automated fact-checking, the broader explainable NLP literature on related social media content moderation tasks has developed feature-level, rationale-based evaluation that informs our approach. In hate speech detection, `HateXplain` (MATHEW et al., 2021) is the first benchmark with human-annotated rationales, and established that models with strong classification performance do not necessarily score well on explanation *plausibility* and *faithfulness*, motivating the evaluation of explanation quality as an axis distinct from accuracy. `HateBRXplain` (SALLES; VARGAS; BENEVENUTO, 2025) extends this paradigm to Brazilian Portuguese, assessing post-hoc attributions (LIME and SHAP) against human rationales and likewise finding that the best-performing classifiers do not always produce faithful rationales. Closer to the modelling side, Eilertsen et al. (2026) proposed Supervised Rational Attention, which aligns model attention with human rationales to improve faithfulness, a self-explaining alternative to the post-hoc attribution methods we adopt. The study of socially-responsible and explainable methods for misinformation and hate speech detection, particularly in lower-resource languages, is further explored by Vargas, Benevenuto, and Pardo (2026). These works reinforce that feature-level attributions must be evaluated for faithfulness rather than assumed, which is the perspective we adopt for the multimodal explanations of mu2X.

5.3 Method

This section introduces mu2X, a multimodal and multilingual framework that provides an end-to-end solution for *explainable* misinformation detection. mu2X extends the foundational mu2 architecture, previously discussed in Chapter 3, by prioritising the extraction and elucidation of key features that explain the classification results, thereby facilitating enhanced human comprehension.

We begin by extending the problem statement from Section 3.3.1, integrating the explainable dimension into the misinformation detection task. Subsequently, we provide the high-level overview of the Classification Explainability Module (Section 5.3.2) of mu2X, followed by a detailed characterisation of their technical implementation.

Whilst mu2X is conceptualised as a holistic end-to-end framework for explainable misinformation detection, the scope of this chapter is specialised towards the formalisation of the explainability methodology used for the detection results. Specific details regarding the underlying classification and detection capabilities are deferred to Chapter 3.

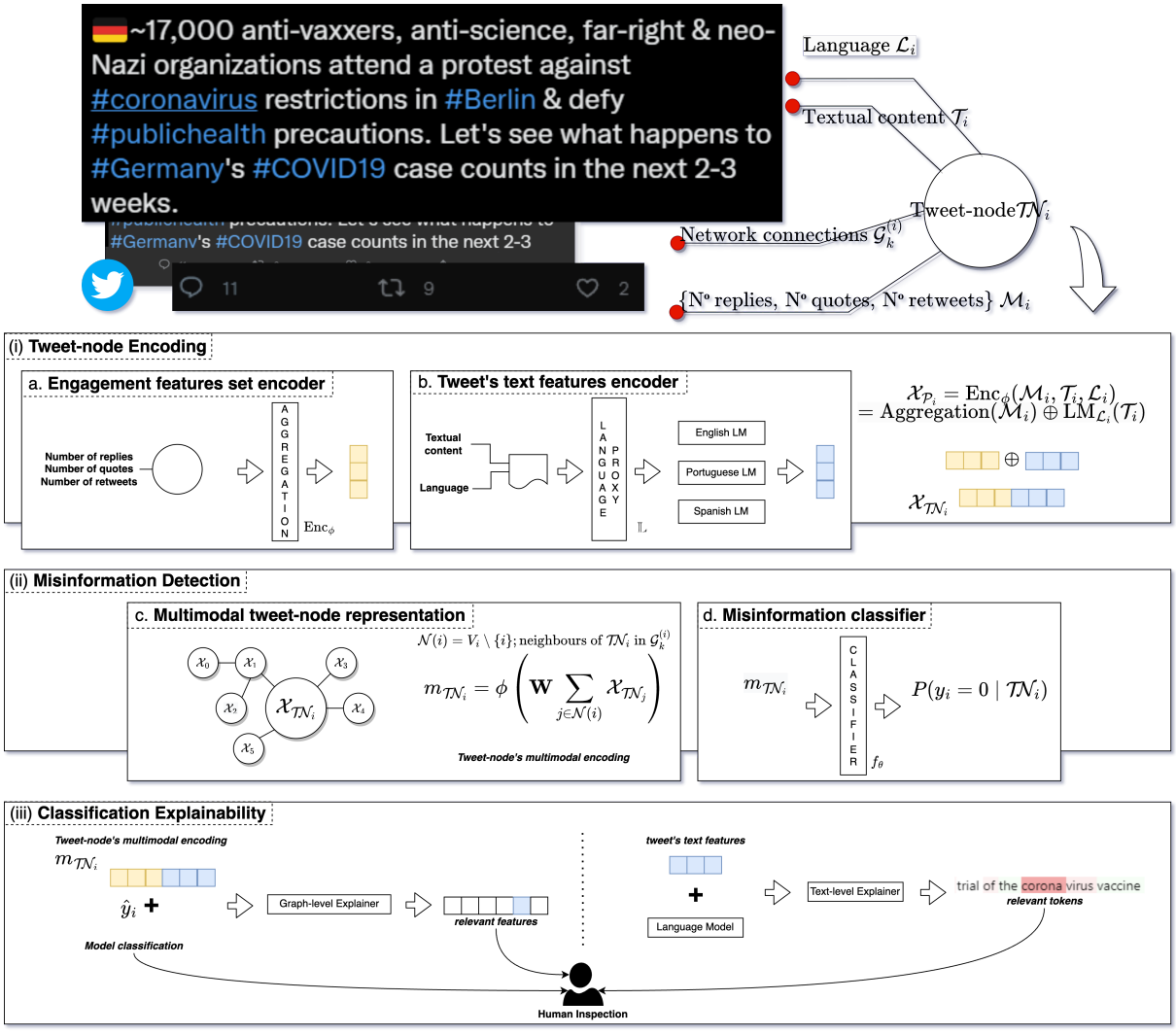


Figure 9: Architectural overview of the proposed mu2X framework instantiated for the Twitter/X ecosystem. Yellow and blue elements represent graph-based and text-based feature vectors, respectively.

5.3.1 Problem Statement

Given a post-node $\mathcal{P}_i = (\mathcal{T}_i, \mathcal{G}_k^{(i)}, \mathcal{M}_i, \mathcal{L}_i)$ as formalised in Section 3.3.1, and a trained classifier $f_\theta : \mathcal{P} \to [0, 1]$ from Section 3.3.3, the objective is to construct, for each prediction $\hat{y}_i = f_\theta(\mathcal{P}_i)$, a two-stage explanation that identifies the most representative features driving the classifier's decision. A graph-based explainer $\zeta_i^{(g)}$ first attributes the decision to the feature dimensions of the multimodal feature representation $\mathcal{X}_{\mathcal{P}_i}$ (Equation (3.2)), which fuses engagement metadata and textual embedding components. When the implicated dimensions correspond to the textual component, whose individual embedding dimensions are not human-interpretable, a token-level explainer $\zeta_i^{(t)}$ refines the explanation by attributing importance to the individual tokens of $\mathcal{T}_i$. By isolating the features that contribute most to the classification of $\mathcal{P}_i$ as *misinformation* or *fact*, we aim to support each automated decision with explanations.

5.3.2 mu2X: Multimodal and Multilingual Explainable Framework for Misinformation Detection

As an extension of mu2, mu2X comprises three main modules: post-node encoding and misinformation detection (both described in Section 3.3.2), and a third novel module *(iii)* classification explainability.

Figure 9 illustrates the instantiation of all three modules, *i.e.*, *post-node encoding*, *misinformation detection*, and *classification explainability*, and the underlying relationship between them.

5.3.3 Classification Explainability

In the Classification Explainability Module, mu2X explains the predicted label using a pair of post-hoc, local explanation methods (Section 2.4.3), one operating over the feature dimensions of the multimodal representation and one over the tokens of the post text. Both explainers are functionals of the trained classifier f_θ and the target post-node $\mathcal{P}_i$. The concrete methods used are specified in the framework instantiation (Section 5.3.4).

The multimodal representation $\mathcal{X}_{\mathcal{P}_i} \in \mathbb{R}^d$, as defined in Equation (3.2), is the concatenation of vector representation of engagement features and a textual embedding vector. We write $I_{\text{txt}} \subseteq \{1, \ldots, d\}$ for the set of dimensions contributed by the textual encoder $\text{LM}_{\mathcal{L}_i}(\mathcal{T}_i)$, and $I_{\text{eng}} = \{1, \ldots, d\} \setminus I_{\text{txt}}$ for the engagement metadata dimensions.

The graph-based explainer $\mathcal{E}_g$ assigns an importance score to each dimension of $\mathcal{X}_{\mathcal{P}_i}$:

$$\zeta_i^{(g)} = \mathcal{E}_g(f_\theta, \mathcal{P}_i), \qquad \zeta_i^{(g)} \in \mathbb{R}^d, \tag{5.1}$$

where a higher score indicates greater influence on the prediction $\hat{y}_i = f_\theta(\mathcal{P}_i)$.

The text-based explainer $\mathcal{E}_t$ assigns a signed importance score to each token of the post text $\mathcal{T}_i$:

$$\zeta_i^{(t)} = \mathcal{E}_t(f_\theta, \mathcal{P}_i), \qquad \zeta_i^{(t)} \in \mathbb{R}^{|\mathcal{T}_i|}, \tag{5.2}$$

where the sign distinguishes tokens that support the prediction from those that oppose it.

The two explainers operate as a two-stage explanatory framework. The graph-based score $\zeta_i^{(g)}$ is computed for every prediction, and the text-based score $\zeta_i^{(t)}$ is computed only when the highest-scoring dimension of $\zeta_i^{(g)}$ belongs to the textual embedding block, that is, when $\arg\max_k \zeta_i^{(g)}[k] \in I_{\text{txt}}$. This second stage provides token-level interpretability in the case where the implicated dimensions are themselves not human-interpretable. The resulting explanations, together with the classifier prediction $\hat{y}_i$, are the final output.

5.3.4 Framework Instantiation

Analogous to the conceptual framework extended in the previous section, this section extends its instantiation from Section 3.3.3 by detailing the solutions used in the Classification Explainability Module, as shown in Figure 9.

Graph-based Feature Explanation We instantiate $\mathcal{E}_g$ with GraphLIME (HUANG, Q. et al., 2023), a non-linear and model-agnostic explainer that extends LIME (RIBEIRO; SINGH; GUESTRIN, 2016) to graph neural networks (Section 2.5.1). To explain the prediction at the target node $\mathcal{P}_i$, GraphLIME fits a sparse surrogate to the classifier's predictions over the k-hop neighbourhood V_i of $\mathcal{P}_i$, using the post-nodes' multimodal representations $\{\mathcal{X}_{\mathcal{P}_j} : j \in V_i\}$ as the regression inputs. The surrogate is fitted with the Hilbert-Schmidt Independence Criterion (HSIC) Lasso (YAMADA et al., 2014), whose non-negative coefficients $\zeta_i^{(g)} \in \mathbb{R}^d_{\geq 0}$ rank the feature dimensions of $\mathcal{X}_{\mathcal{P}_i}$ by their relevance to the prediction.

Text-based Feature Explanation We instantiate $\mathcal{E}_t$ with Integrated Gradients (IG) (SUNDARARAJAN; TALY; YAN, 2017), applied whenever the highest-scoring dimension of $\zeta_i^{(g)}$ belongs to the textual embedding block I_{txt}. IG is a gradient-based attribution method

that assigns an importance score to each input token by accumulating the gradients of the classifier output f_θ along a straight-line path from a baseline[1] to the actual input. By construction, the tokens' scores sum to the difference between the prediction and the baseline output. We normalise them to $[-1, 1]$ for presentation, translating the coarse textual embedding attribution into token-level scores.

5.4 Experimental Setup

Trustworthiness Evaluation Protocol The trustworthiness evaluation protocol assesses whether an explanation exposes the features that the prediction actually depends on, which is the information a user needs to decide when to trust a prediction. A prediction that relies on features a user would reject is only detectable as such when the explanation reveals that reliance. The protocol adapts the simulated user trust experiment (RIBEIRO; SINGH; GUESTRIN, 2016) to the multimodal post-node setting and uses the notion of forward-simulatability (DOSHI-VELEZ; KIM, 2017), which establishes that a useful explanation should enable a user to reach the same trust judgment of an oracle with full feature access. Figure 10 depicts the four-stage pipeline. All stages use the trained classifier f_θ of Chapter 3 without retraining, and feature removal is performed by zero-masking the corresponding dimensions of the multimodal representation.

In the first stage, *Feature Simulation*, we randomly select ρ[2] of the features of each instance x and designate these as "untrustworthy", yielding an untrustworthy set F_{untrust} that serves as a controlled proxy for the features a user would distrust.

In the second stage, *Oracle Construction*, the oracle is a labelling function with full access to F_{untrust}. All untrustworthy features are zero-masked in x to produce x', which is reclassified by f_θ. The oracle establishes the ground truth label by marking the prediction *untrustworthy* if the label changes ($f_\theta(x) \neq f_\theta(x')$), since the designated features are then decisive, and *trustworthy* otherwise.

In the third stage, *Simulated User Judgment*, the surrogate g produces an explanation $g(x)$, given by its top-K features by importance. We simulate user judgment by zero-masking only the untrustworthy features from x that appear in $g(x)$, producing x'', which is reclassified by f_θ. The simulated user labels the prediction *untrustworthy* if the label

[1] We use the default baseline of the Captum implementation (KOKHLIKYAN et al., 2020), a zero-valued tensor of the same shape as the input. [2] The experiment described in Section 5.5.2 is averaged over repeated draws, where we use $\rho = 0.30$ across 25 rounds.

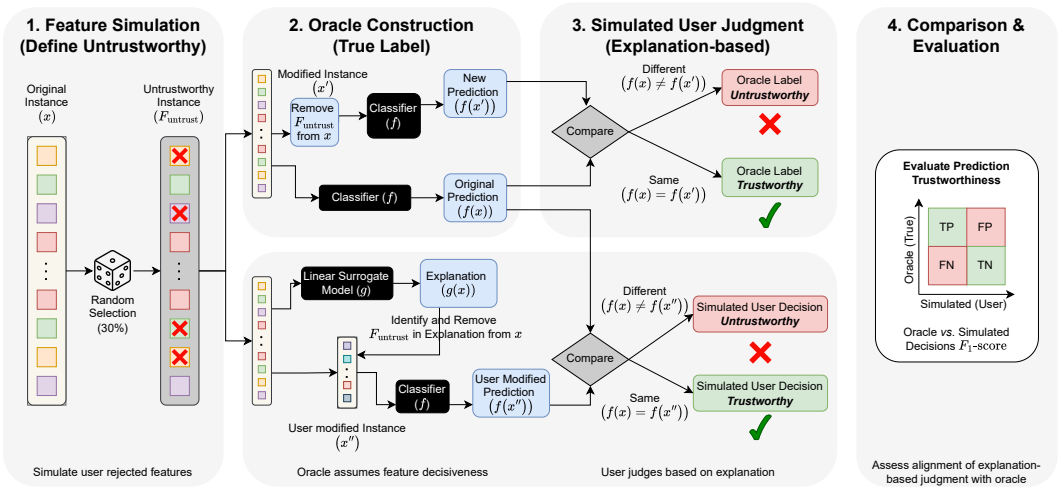


Figure 10: Trustworthiness Evaluation Protocol pipeline, comprising four stages: Feature Simulation, Oracle Construction, Simulated User Judgment, and Comparison and Evaluation.

changes ($f_\theta(x) \neq f_\theta(x'')$), and *trustworthy* otherwise.

Finally, in the fourth stage, *Comparison and Evaluation*, the oracle label is compared with the simulated user label across instances, and the F_1-score is computed with *trustworthy* as the positive class to quantify their agreement. A high F_1 indicates that the judgment supported by the explanation recovers the oracle's verdict often enough to reliably tell trustworthy predictions apart from untrustworthy ones.

Robustness of Explanations Evaluation Protocol We further propose a three-stage protocol, illustrated in Figure 11, to evaluate the capacity of the explainable method to identify relevant, *i.e.*, non-noisy features. The proposed protocol shares its motivation with established work on explanation robustness (ALVAREZ-MELIS; JAAKKOLA, 2018) and with retraining-based interpretability benchmarks such as the remove-and-retrain procedure (HOOKER et al., 2019). The underlying principle, that a faithful explainer should not attribute importance to features uncorrelated with the label, is the one tested by the sanity checks (ADEBAYO et al., 2018). Our variant inverts the remove-and-retrain design, where we *inject* noisy features and measure how often the explainer wrongly selects them instead of removing the most important features.

In the first stage, *Feature Simulation*, $N = |\mathcal{X}| \times p$ *noisy features* are artificially introduced into each sample $\mathcal{X}$, where $N \in \mathbb{N}$, $|\mathcal{X}|$ is the number of original features in the sample, and $p \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1.0\}$ denotes the proportion of noisy features

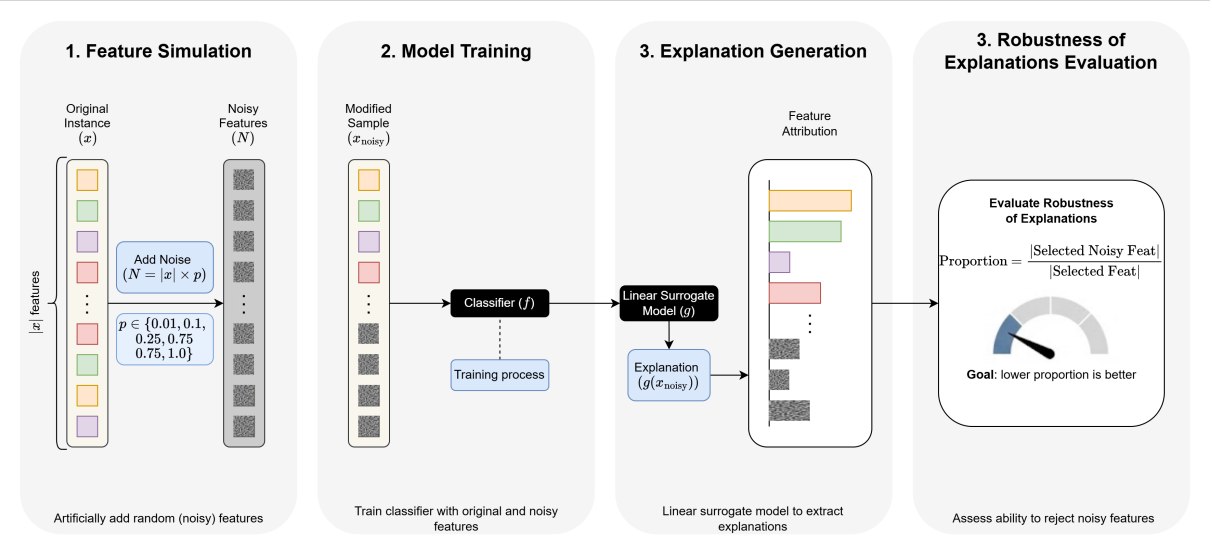


Figure 11: Robustness of Explanations Evaluation Protocol pipeline, comprising three stages: Feature Simulation, Model Training, and Explanation Generation and Evaluation.

added. Each noisy feature is drawn from a uniform distribution on the same scale as the original features and independently of the label, so that any importance the explainer assigns to it is by construction an error. In the second stage, *Model Training*, the classifier is retrained on the augmented samples. Finally, in the third stage, *Explanation Generation and Evaluation*, the surrogate explainer extracts feature attributions from the retrained classifier, and the proportion of noisy features selected as explanatory factors is measured, where a lower proportion indicates greater robustness.

5.5 Results and Analysis

We examined the explanatory features through a combination of qualitative and quantitative analyses.

5.5.1 Interpretability of Explanations

Our qualitative examination focuses on the most relevant features and their ability to connect the tweet-node and its corresponding veracity label. The outcomes of these investigations are detailed in Table 8, which presents the textual content and metadata alongside the GAT model's classification, the discriminative features identified by GraphLime (HUANG, Q. et al., 2023), and our formal interpretation of these explanatory factors.

In Table 8, the first entry is classified as misinformation by the GAT model, with

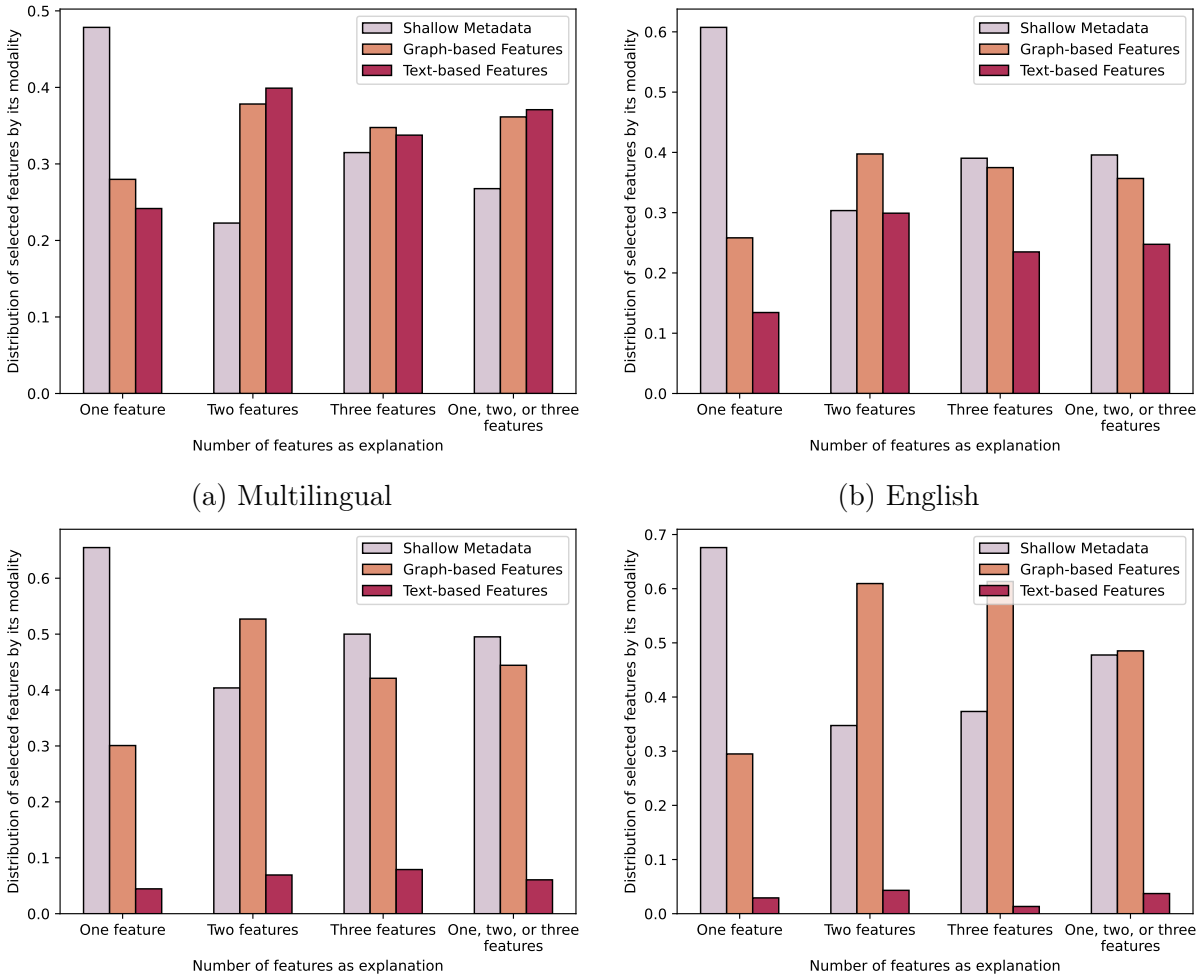
Table 8: Report of two inferred cases with their classification, features, and explanatory factors identified by our framework.

Text	Shallow metadata set	Classification	GraphLime most representative features	Human interpretation
Great news! Carona virus vaccine ready. Able to cure patient within 3 hours after injection. Hats off to US Scientists. Right now Trump announced that Roche Medical Company will launch the vaccine next Sunday, and millions of doses are ready from it !!! VIA: @wajih79273180 https://t.co/BZJCLtwuXq	Number of retweets 26 Number of replies 42 Number of quotes 7	Misinformation	Number of retweets Number of replies	Users who retweet tend to spread misinformation
17,000 anti-vaxxers, anti-science, far-right & neo-Nazi organizations attend a protest against #coronavirus restrictions in #Berlin & defy #publichealth precautions. Let's see what happens to #Germany's #COVID19 case counts in the next 2-3 weeks. https://t.co/TD5xIoT5sV	Number of retweets 11 Number of replies 9 Number of quotes 2	Fact	Text	Word importance illustrated in Fig. 12.

Figure 12: Word importance for text explainability of the second tweet from the table above.

GraphLime isolating the retweet and reply counts as the most relevant features for this decision. Upon analysing these interactions, we observe that users engaged, *i.e.*, those who retweet or reply to these posts, frequently have other tweets classified as misinformation, suggesting a structural trend in the dissemination of misinformation. The subsequent entry was classified as factual, where textual embeddings emerged as the most feature explanatory factor. Furthermore, we investigated textual factors using Captum (KOKHLIKYAN et al., 2020) to interpret the linguistic drivers of the classification label (Figure 12). Notably, phrases such as "protest against #coronavirus" and "Let's see what happens to #Germany" contributed directly to the accurate classification.

For our quantitative analysis, we used a fixed multimodal setting to study how language and dataset size influence these explanatory factors. Figure 13 shows, for each language (English, Spanish, Portuguese, and the aggregation – multilingual), the distribution of features selected as explanations by modality (metadata, graph-based, or text-based) and by the number of features chosen (one, two, three, and the aggregation). In lower-resource linguistic contexts, such as Spanish (Figure 13.c) and Portuguese (Figure 13.d), metadata and graph-based features exhibit higher utility, followed by textual features. This phenomenon is likely attributable to the reduced node density and sparser relational neighbourhood, which renders these structural features more discriminative. Conversely, the multilingual aggregate (Figure 13.a) prioritises textual features, likely due to the increased complexity and density of the associated network. The English subset (Figure 13.b) maintains a balanced distribution across modalities, reflecting its intermediate intermediate density. Irrespective of the language, metadata features remain the most frequent explanatory factor in single-feature selections, reinforcing the premise that interaction metrics serve as



(a) Multilingual

(b) English

(c) Spanish

(d) Portuguese

Figure 13: Distribution of features selected as explanations by their modality, with results separated by language and colour indicating the feature sets.

robust indicators of misinformation bubbles, as seen in the first entry of Table 8.

5.5.2 Trustworthiness of Explanations

In real-world applications, it is fundamental that the classifier's predictions are credible, *i.e.*, the classifier should be consistently accurate in associating labels so that the end-user can relate to the prediction as trustworthy. Moreover, explanations play an essential role in assessing predictions. Therefore, we develop a protocol to measure the extent to which the explanations have the capacity to assist users in determining the reliability of the prediction.

Given the trustworthiness evaluation protocol, we report and compare the F1-score on the trustworthy predictions for the classifier trained with graph-based features, text-based

Table 9: F1-Score of mu2X's trustworthiness based on different modalities.

Modality	F1-Score given top-k most relevant features				
	Top-1	Top-2	Top-3	Top-5	Top-10
Graph-based	0.7576 ± 0.2199	0.8059 ± 0.0158	0.8035 ± 0.0188	**0.8076 ± 0.0137**	*0.8031 ± 0.0187*
Text-based	*0.8188 ± 0.0164*	**0.8096 ± 0.0166**	**0.8111 ± 0.0154**	0.8060 ± 0.0156	0.8029 ± 0.0118
Multimodal	**0.8207 ± 0.0185**	*0.8061 ± 0.0153*	*0.8080 ± 0.0203*	*0.8062 ± 0.0167*	**0.8115 ± 0.0175**

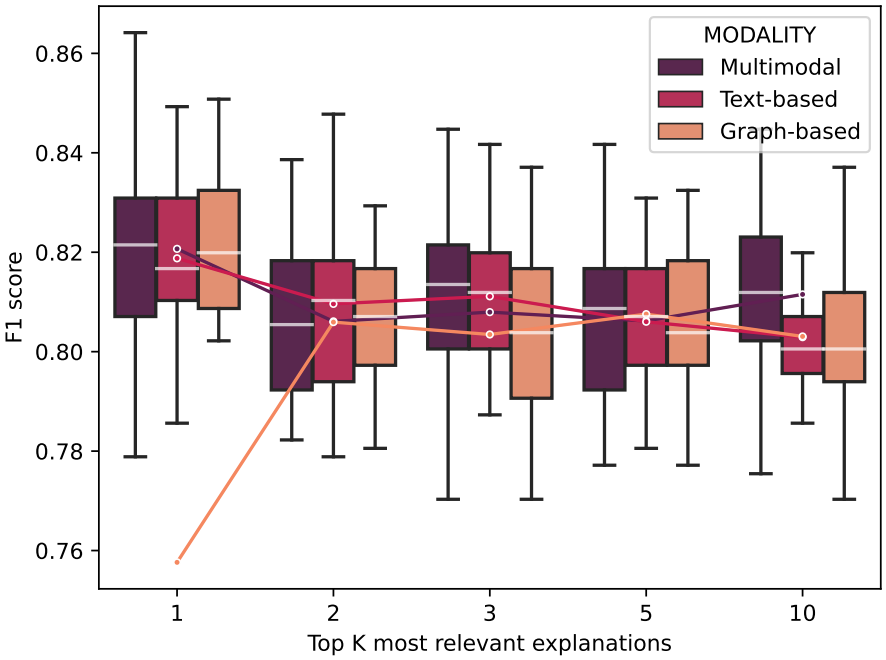


Figure 14: F1-Score of trustworthiness box-plot visualisation. Lines denote the average F1-Score for each Top-K.

features, and multimodal features over 25 rounds, considering the top-K most relevant explainable features ($K \in \{1, 2, 3, 5, 10\}$). The results are compiled in Fig. 14 and Table 9. The multimodal classifier achieved the most stable results, having the best F1-score in top-1 and top-10, whilst being the second-best in the other cases. The classifier that was solely trained with graph-based features displayed the worst results, indicating that it usually mistrusts trustworthy predictions, whilst trusting untrustworthy predictions too often. Despite the used modality, the results show F1-Score above 80% (except the graph-based top-1), confirming the reliability of the mu2 explanations.

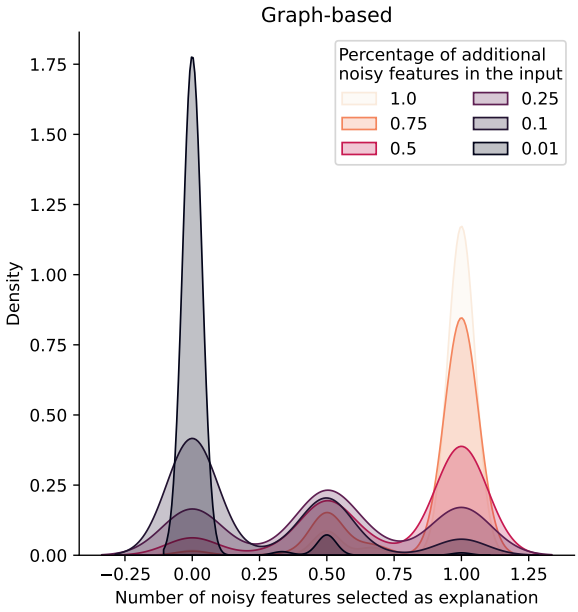
5.5.3 Robustness of Explanations

Besides credible predictions, robust explanations are key for the success of misinformation classifiers. Given the robustness of explanations evaluation protocol, we report the average results of 25 rounds of the noisy features selected as explanation patterns for the three classifiers. In Figure 15, we show the kernel density estimation of the frequency of selected noisy features given the classifier and the proportion of noisy features added. The explanations based on the text-based features classifier were more robust, having a high density at zero noisy features selected in the lower percentages and even curves at zero or one noisy feature selected in higher noisy environments. The explanations grounded on the multimodal features classifier showed average results, where it keeps the higher density at zero for lower noisy environments and is even out on higher noisy environments.

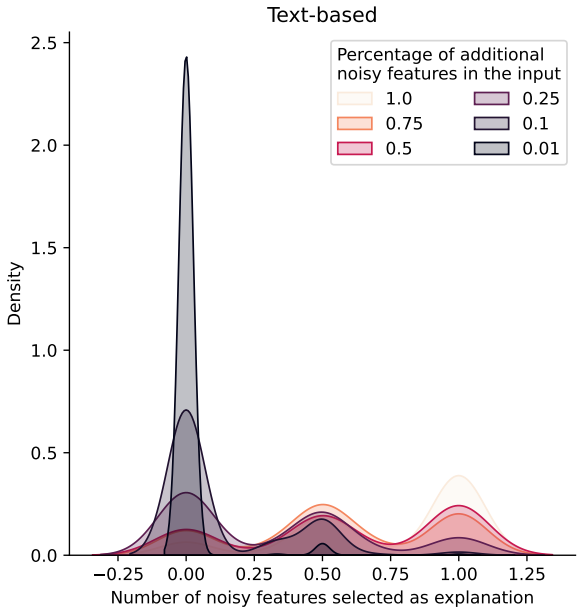
These patterns can also be observed in Fig. 15d. The figure shows the distribution between the average percentage of noisy features selected as explanations and the percentage of noisy features added. Here, it is observed that the features selected as explanations for the text-based classifier increase linearly with the percentage of noisy features added. Additionally, it can be observed that the features selected as explanations using the multimodal classifier soften the non-linear curve presented by the explanations over the graph-based classifier.

5.6 Conclusion and Discussions

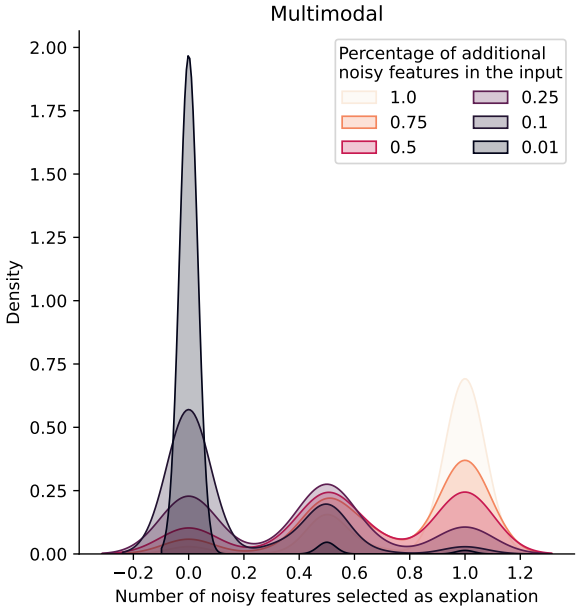
In this chapter, we proposed mu2X, a multimodal and multilingual explainable framework for misinformation detection that extends mu2 (Chapter 3) with a Classification Explainability Module combining GraphLIME for relational features and Integrated Gradients for textual features. Our experiments on the `MuMiN` dataset across English, Portuguese, and



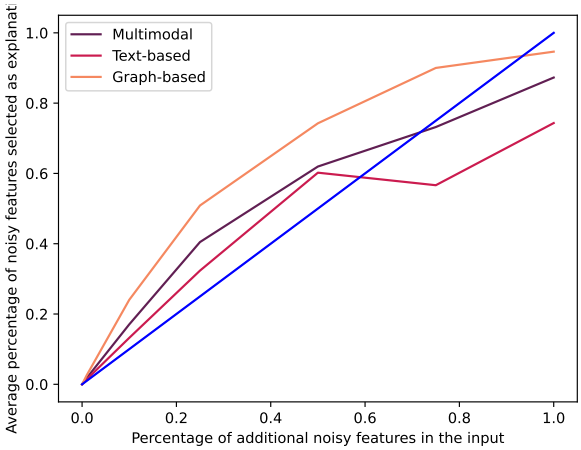
(a) Graph-based features



(b) Text-based features



(c) Multimodal features



(d) Relation between the percentage of additional noisy features and the average percentage of noisy features selected as an explainable factor.

Figure 15: Distribution of features selected as explanations by their modality, with results separated by language and colour indicating the feature sets.

Spanish demonstrate that the proposed explainability methodology adds completeness to the classification framework by providing trustworthy, robust, and interpretable feature-level explanations. The trustworthiness evaluation reveals that multimodal explanations are the most stable across top-K feature selections, achieving the best F1-scores at top-1 (0.8207 ± 0.0185) and top-10 (0.8115 ± 0.0175), with all modalities exceeding 0.80 F1 except graph-based at top-1 (0.7576 ± 0.2199). The robustness evaluation shows that text-based explanations exhibit the greatest resilience to noisy feature injection, with a desirable linear increase in noisy feature selection, whilst graph-based explanations are the least robust, displaying a non-linear response; multimodal explanations occupy an intermediate position, softening the non-linear behaviour of graph-based features. Our interpretability analysis reveals that metadata features (interaction counts) are the most frequent single-feature explanation across all languages, reinforcing that engagement metrics serve as robust indicators of misinformation dissemination patterns. Notably, low-resource languages rely more heavily on metadata and graph-based features due to sparser relational neighbourhoods. These findings directly address Research Question 3, demonstrating that multimodal feature integration enhances both the interpretability and the trustworthiness of explanations, whilst text-based features provide the most robust explanatory signal.

There are several promising directions for future work to build upon these findings. First, we aim to enhance mu2X by incorporating additional modalities (such as image and video content) alongside modality-specific explainability methods; we hypothesise that leveraging richer data sources will improve both classifier performance and the completeness of the generated explanations. Second, the development of a protocol for quantitatively evaluating interpretability based on human assessment represents a critical gap, since existing evaluation methodologies do not adequately handle multimodal explanations; such a protocol would complement the trustworthiness and robustness evaluations proposed in this chapter. Third, we envision integrated explanations that combine content snippets with explanatory factors or incorporate prior knowledge to improve clarity and user understanding, moving beyond feature-level attributions towards more natural and actionable justifications. Fourth, the application of the proposed trustworthiness and robustness evaluation protocols to other explainable AI systems beyond misinformation detection would validate their generalisability and contribute to the broader XAI evaluation literature. Finally, extending the evaluation to additional datasets and languages would strengthen the generalisability of our findings, particularly for typologically diverse and lower-resource languages where the balance between modalities

may differ substantially. The integration of feature-level explainability with LLM-based natural language justifications, combining attributions with human-readable reasoning, is further explored in Chapter 6.

5.7 Limitations

The findings reported in this chapter are subject to two categories of limitations, those inherited from feature-level explainability for multimodal classification and those specific to the design and scope of mu2X.

The most impactful general limitation is *user interpretability*. Feature-level explanations such as importance scores over metadata attributes or word-level attributions can be difficult for non-expert users to interpret. They provide precise and complete information about model behaviour but require a degree of technical literacy that may limit their practical utility for content moderators or end users, reflecting the interpretability-completeness trade-off discussed in Section 2.4.2. Closely related is the issue of *post-hoc approximation*. GraphLIME and Integrated Gradients both approximate the behaviour of the underlying classifier rather than providing inherently interpretable decisions, and the fidelity of these approximations depends on the complexity of the model and the local decision boundary. Thus, they may not fully capture the reasoning process in all cases. A third concern is the open status of *explanation-quality evaluation* in the XAI literature. The trustworthiness and robustness protocols introduced here provide quantitative assessments, but they rely on simulated user behaviour rather than on actual human evaluation, and may not fully capture the practical utility of the generated explanations.

Other limitations worth noting fall into two groups. The first concerns the *methodological scope of the evaluation*. Consistent with the classification framework of Chapter 3, we evaluate only vector concatenation as the multimodal feature-aggregation strategy, and the trustworthiness protocol of Section 5.5.2 uses a linear surrogate as a simulated user rather than actual human assessments. Both choices favour reproducibility and tractable comparison across the three languages of `MuMiN`, since exhaustively comparing fusion mechanisms or recruiting human annotators at scale would dilute the analysis of the explainability dimension we set out to investigate. Alternative fusion strategies and a calibrated human-evaluation study are deferred to future work. The second concerns *coverage*. The evaluation is conducted exclusively on the `MuMiN` dataset across three languages (English, Portuguese, and Spanish), inheriting the same scope rationale as Chapter 3.

MuMiN is among the few multilingual benchmarks that combine social-network metadata with the modalities required by mu2X, and its annotated content drives the three-language scope. Validation on additional benchmarks with different annotation schemes, network structures, and typologically diverse languages would strengthen the generalisability of both the framework and the proposed evaluation protocols.

6 Explainable Knowledge Graph-based Contrastive Reasoning with LLMs for Misinformation Detection

This chapter presents the explainability dimension of KG-CRAFT, the framework introduced in Chapter 4, augmenting it with a dedicated justification generation step that consumes the contrastive summary $A_{\mathcal{C}}$ produced by KG-CRAFT's contrastive reasoning component (Section 4.3.3). The chapter examines the natural language justifications produced by this extension. The veracity assessment performance of KG-CRAFT is addressed in Chapter 4.

6.1 Introduction and Background

Whilst KG-CRAFT (Chapter 4) achieves state-of-the-art claim verification performance for automated fact-checking (Section 2.3), the use of large language models (LLMs) (Section 2.6) as the main mechanism for verification in fact-checking raises concerns related to the opacity of the underlying decisions and the capability of producing fluent yet factually inconsistent outputs (WANG, Yuxia et al., 2024; AUGENSTEIN et al., 2024). As discussed in Section 2.4, explainable methods seek to address this challenge by providing mechanisms that render model decisions transparent and understandable to diverse stakeholders. In the context of automated fact-checking and the different users' perspectives (Section 2.4.1), the need for explainability requires not only accurate veracity labels but also interpretable justifications for why a given claim is assessed as a certain label (*e.g.*, true or false) (KOTONYA; TONI, 2020a; SAEED; OMLIN, 2023). Without such justifications, automated systems risk undermining the very trust they are designed to protect (GILPIN et al., 2018; WARREN; SHKLOVSKI; AUGENSTEIN, 2025).

Among the explanation paradigms discussed in Section 2.4, we focus on natural language justifications, *i.e.*, free-text rationales generated alongside the veracity label that articulate, in human-readable form, the reasoning supporting the assessment. This focus

is motivated by the operational characteristics of LLM-based verifiers, whose reasoning is itself produced as natural language and is therefore most faithfully exposed through textual justifications rather than through feature-level attributions of the kind explored in Chapter 5. Existing approaches to explainable LLM-based fact-checking, however, have predominantly framed justification generation as a side-task evaluated either through small-scale human studies, which face inherent generalisability limitations (KOTONYA; TONI, 2020a), or through reference-based metrics such as ROUGE and BLEU that conflate surface-level lexical overlap with explanatory quality (ATANASOVA et al., 2020; YAO et al., 2023; RUSSO; TEKIROĞLU; GUERINI, 2023). A major limitation of these methods, however, is their failure to assess the multiple, complementary qualities that an explanation must possess to be useful for downstream decision-making. We argue that a comprehensive evaluation of LLM-generated justifications must assess several explanation desiderata jointly (DOSHI-VELEZ; KIM, 2017; GILPIN et al., 2018; KOTONYA; TONI, 2020b; MINH et al., 2022), examining justifications' fluency and whether they are actionable, causally structured, contextually grounded, impartial, parsimonious, and safe with respect to the misinformation they purport to refute.

In this chapter, we empirically investigate how the structured contrastive reasoning that drives KG-CRAFT's veracity-assessment performance translates into the quality of its natural language justifications. We propose a multi-desiderata evaluation protocol that operationalises six standard XAI desiderata via LLM-as-a-Judge prompts on a Likert scale, and apply it to four ablation configurations of KG-CRAFT alongside human-written reference explanations on the `LIAR-RAW` (YANG, Z. et al., 2022) test set. This chapter addresses Research Question 4, which asks how the choices of evidence representation (structured *vs.* unstructured) and reasoning formulation (contrastive *vs.* direct) shape the quality of LLM-generated justifications across complementary explainability desiderata. The main contributions of this chapter are as follows:

- We introduce the explainability extension of KG-CRAFT, which augments the framework of Chapter 4 with a justification generation component. The component reuses KG-CRAFT's contrastive summary $A_{\mathcal{C}}$ and, via a dedicated prompt p_{just}, produces a natural language justification for a given verdict $\mathcal{V}_{\mathcal{C}}$;
- We present a multi-desiderata evaluation protocol for LLM-generated fact-checking justifications, operationalising six standard XAI desiderata (*Actionable*, *Causal*, *Contextful*, *Impartial*, *Parsimonious*, *Safety*) via LLM-as-a-Judge prompts on a 5-point Likert scale, with five-run aggregation for stochastic robustness;

- We conduct a systematic explanation-quality ablation of KG-CRAFT's components, paralleling the verification-accuracy ablations of Sections 4.5.2 and 4.6.1, and demonstrate that KG-CRAFT's justifications outperform every ablation on five of the six desiderata, reaching approximately 87.5% of the reference explanation quality on average and exhibiting their largest improvement on the *Impartial* dimension (+0.20 Likert points over the next-best ablation);
- We provide a structured failure-mode analysis of disagreement cases, characterising the complementary failure profiles of the human-written reference explanations and KG-CRAFT-generated justifications; and
- We complement the LLM-as-a-Judge results with reference-based metrics (ROUGE, BLEU, BERTScore), showing that the human-written reference explanations diverge more from the source evidence than LLM-generated justifications do, so that lexical-overlap metrics would contradict rather than corroborate the judge's assessment of justification quality, validating the methodological choice of LLM-as-a-Judge evaluation.

In the following sections, we first present the related work on explainable LLM-based fact-checking and on LLM-as-a-Judge evaluation (Section 6.2). We then present the justification generation extension of KG-CRAFT, detailing the problem statement, the six desiderata, and the LLM-as-a-Judge scoring procedure (Section 6.3). Subsequently, we describe the experimental setup (Section 6.4), report and analyse the results across desiderata, per-claim alignment with the reference, reference-based textual alignment, and qualitative failure modes (Section 6.5), and conclude with a discussion of findings and future directions (Section 6.6).

6.2 Related Work

This section surveys the literature most relevant to the contributions of this chapter, organised into three parts: *(i)* the desiderata-based characterisation of explanations for fact-checking, which supports our choice of evaluation dimensions; *(ii)* the wider context of explainable LLM-based fact-checking, which situates the proposed evaluation within current methodological practice; and *(iii)* LLM-as-a-Judge evaluation, which motivates the methodological choice of automated multi-desiderata scoring.

6.2.1 Desiderata for Fact-Checking

The systematic characterisation of what constitutes a good explanation in the automated fact-checking domain was pioneered by (KOTONYA; TONI, 2020b), who, building on the explainable artificial intelligence (Section 2.4) literature (DOSHI-VELEZ; KIM, 2017; GILPIN et al., 2018; LIPTON, Z. C., 2018), proposed a set of intelligibility properties that fact-checking explanations should satisfy: *actionable*, *causal*, *coherent*, *contextful*, *interactive*, *impartial*, *parsimonious*, and *chronological*. Their analysis surveyed existing explainable automated fact-checking systems with respect to these properties and highlighted the persistent trade-off between explanation *interpretability* and explanation *completeness* (Section 2.4.2). Therefore, it is argued that a comprehensive evaluation should consider multiple complementary properties simultaneously rather than collapsing them into a single scalar (LIPTON, Z. C., 2018; NAUTA et al., 2023).

Subsequent work has applied these properties as evaluation dimensions, predominantly through human-centric studies that constrain the assessment to a single property at a time and to small claim samples (KOTONYA; TONI, 2020a; YAO et al., 2023). The protocol introduced in this chapter (Section 6.3) operationalises five of the eight Kotonya-Toni properties at scale via LLM-as-a-Judge. We retain *Actionable*, *Causal*, *Contextful*, *Impartial*, and *Parsimonious*, which can be assessed from the justification text alone, and additionally introduce a *Safety* desideratum that captures the misinformation-amplification risk specific to fact-checking deployment (Section 2.3) (WARREN; SHKLOVSKI; AUGENSTEIN, 2025). We exclude *Coherent*, since incoherent text would fail to admit a meaningful Likert score under any of the remaining desiderata; *Interactive*, which presupposes a human-in-the-loop evaluation incompatible with batch protocols; and *Chronological*, which is a dataset-level property of the underlying ruling reports rather than of the justification text and is not informative on the snapshot evaluations conducted in this chapter.

6.2.2 Explainability of LLM-based Fact-Checking

We now situate the proposed evaluation within the wider landscape of explainable fact-checking, focusing on systems that produce natural language justifications, the explanation form addressed in this chapter, whose artefacts are candidates for the multi-desiderata assessment introduced above. Early approaches focused on extractive justifications, identifying salient sentences from the supporting evidence as the basis for the veracity decision. `GenFE` and `GenFE-MT` were both proposed by (ATANASOVA et al., 2020), which jointly optimise veracity prediction and explanation generation through multi-task learning,

demonstrating that the two objectives are mutually reinforcing. Building on this foundation, (KOTONYA; TONI, 2020a) tailored a domain-specific framework for the automated fact-checking of public health claims, combining extractive evidence selection with abstractive summarisation. The authors conducted one of the first human-centric evaluations of generated explanations, highlighting the importance of evaluating not only what is generated but also how it is perceived by domain experts. We note, however, that such human evaluations, whilst valuable, are difficult to scale, relying, typically, on small samples and few annotators, which limits the generalisability of their findings and motivates complementary, automated evaluation at scale.

The emergence of LLMs has enabled a paradigm shift towards generative justifications that integrate evidence retrieval, reasoning, and explanation in a unified pipeline (ELDIFRAWI; WANG; TRABELSI, 2024). Guo, Zeng, et al. (2023) proposed `IKA`, which constructs a graph of positive and negative examples from labelled data to enhance both fact-checking and explanation capabilities. Hui Liu et al. (2024) introduced `TELLER`, a dual-system framework that emphasises trustworthiness through the integration of human expertise and LLM capabilities. Bo Wang et al. (2024) developed `L-Defense`, a defence-based framework that addresses the limitations of uncensored crowd wisdom by splitting evidence into competing narratives and leveraging LLMs for reasoned verification. Yao et al. (2023) introduced `Mocheg`, a multimodal, multi-topical dataset of articles, along with an end-to-end framework that generates ruling statements by leveraging the input claims, the predicted veracity label, and supporting textual evidence. Jinguang Wang et al. (2025) proposed `EExpFND`, an end-to-end framework that jointly performs fake news detection and explanation generation through evidence-claim variational causal inference, modelling the two tasks together rather than separately. More recently, Vargas, Salles, et al. (2024) proposed `SELFAR`, which performs sentence-level factual reasoning and generates post-hoc explanations via attribution methods and zero-shot LLM prompts, reporting that zero-shot prompts yield readable explanations but remain prone to hallucination, which reinforces the need for the rigorous justification evaluation we pursue. Extending justification generation towards full fact-checking articles, `QRAFT` (SAHNAN et al., 2026) proposes an LLM-based agentic framework that mimics the workflow of human fact-checkers, evaluated against expert-written articles, which it approaches but does not yet match.

We note that, whilst these approaches collectively advance the generation of natural language justifications, the evaluation methodologies adopted in the literature remain fragmented. Reference-based metrics, such as ROUGE and BLEU, are inherited from the broader text-generation community and evaluate lexical overlap with reference texts.

Such metrics are poorly suited to explanation evaluation because lexical overlap can be high even when the explanation is logically inconsistent with the assigned label, and conversely, abstractive yet faithful explanations may exhibit low overlap. Furthermore, the human evaluations typically used to complement these metrics are constrained in scale and may not generalise across diverse claim populations. This chapter addresses this gap by combining the desiderata catalogue (KOTONYA; TONI, 2020b) with an automated, scalable scoring procedure based on LLM-as-a-Judge, applied to the natural language justifications produced by KG-CRAFT and its ablations.

6.2.3 LLM-as-a-Judge for Explanation Evaluation

A growing number of recent works have investigated the use of large language models as automatic evaluators of generated text, leveraging their capacity for linguistic judgment to provide scalable and reproducible quality assessments (ZHENG et al., 2023; LI, H. et al., 2024; GU et al., 2026). This paradigm, commonly referred to as LLM-as-a-Judge, has been applied to a wide range of generation tasks (CHANG et al., 2024; LI, D. et al., 2025), including dialogue evaluation (LIU, Y. et al., 2023; FU et al., 2024), summarisation quality assessment (LIU, Y. et al., 2023; GAO, M. et al., 2025), and instruction-following benchmarking (ZHENG et al., 2023), and has been extended to fine-grained, criteria-based scoring against explicit rubrics (KIM, S. et al., 2024; YE et al., 2024). Additional reference-free metrics for information alignment include *AlignScore* (ZHA et al., 2023), which performs automatic factual consistency evaluation of text pairs, and *RQUGE* (MOHAMMADSHAHI et al., 2023), which evaluates generated questions based on the corresponding context and answer. Both metrics are employed for fact-checking specifically in the verification-quality analysis of Section 4.5.2.1. A related line of work follows a *decompose-then-verify* paradigm, breaking long-form text into atomic claims and using an LLM to verify each against evidence, as in `FActScore` (MIN et al., 2023), `SAFE` (WEI, Jerry et al., 2024), and `MiniCheck` (TANG; LABAN; DURRETT, 2024). These methods illustrate that LLM-based judgment is already employed for fact-checking sub-tasks, namely assessing the factual support of individual claims against evidence, even though they target factual precision or consistency rather than the broader qualities of an explanation. At the system level, `RealFactBench` (YANG, S. et al., 2025) is a benchmark that uses LLMs to evaluate the fact-checking capabilities of LLMs and multimodal LLMs across knowledge validation, rumour detection, and event verification, confirming that LLM-as-a-Judge is increasingly adopted within the fact-checking domain itself. These metrics, whilst valuable for measuring specific facets of generation quality, target a single dimension each and therefore do not address the

multiple complementary desiderata reviewed in Section 6.2.1. Beyond fact-checking, related work on social-media content moderation has begun to evaluate LLM-generated rationales directly: (TRAGER et al., 2025) introduced `MFTCXplain`, a multilingual benchmark that assesses the moral reasoning of LLMs through multi-hop hate speech explanations and reports a substantial divergence between human- and model-generated rationales, including for Llama 3.3 70B under zero-shot, few-shot, and chain-of-thought prompting. This corroborates, on the same backbone we adopt, the human–model explanation gap that our qualitative failure analysis examines in Section 6.5.4.

Whilst LLM-as-a-Judge protocols offer substantial scalability advantages, the literature has documented several systematic biases that must be considered (LI, D. et al., 2025; GU et al., 2026). These include positional and length biases (ZHENG et al., 2023; WANG, P. et al., 2024), where evaluators favour longer or earlier-presented outputs; self-preference biases (PANICKSSERY; BOWMAN; FENG, 2024; CHEN, Z.-Y. et al., 2025), where evaluators systematically favour outputs from models in the same family; and prompt sensitivity (WANG, P. et al., 2024; POMBAL; REI; MARTINS, 2026), where small wording changes in the evaluation prompt can shift the rank ordering of compared methods. Multi-run aggregation reduces the stochastic variance of LLM judgments, whereas the systematic biases above are more directly addressed by panels of judges from different model families (VERGA et al., 2024). The protocol proposed in this chapter adopts the former (five independent runs aggregated by mean and standard deviation) to control sampling variance, and identifies multi-family panels as future work (Section 6.6). The residual self-preference bias is examined separately in Section 6.7.

Building on these observations, the remainder of this chapter formalises the justification generation extension of KG-CRAFT (Section 6.3) and the multi-desiderata evaluation protocol applied to it (Section 6.4).

6.2.4 Datasets for Explainable Fact-Checking

This section reviews the principal datasets that pair fact-checked claims with annotated justifications, with particular emphasis on those relevant to the evaluation of justification quality in this chapter.

LIAR-RAW (YANG, Z. et al., 2022) Beyond the report-level evidence described in Section 4.2.3.4, `LIAR-RAW` associates each claim with a human-written PolitiFact justification, which serves as the reference explanation against which generated justifications are com-

pared in this chapter.

RAWFC (YANG, Z. et al., 2022) As described in Section 4.2.3.4, `RAWFC` pairs Snopes claims with their retrieved reports; like `LIAR-RAW`, it also provides the fact-checker's justification, making it suitable for evaluating explanation quality alongside verification.

ExClaim (GURRAPU; HUANG; BATARSEH, 2022) Built from 4,006 PolitiFact articles, `ExClaim` pairs each claim with its source, ruling-comment evidence, and a binary verdict (`supports`, `refutes`) collapsed from PolitiFact's six classes. It was designed for explainable claim verification through rationalisation, generating a natural language rationale that justifies the verdict.

HealthFC (VLADIKA; SCHNEIDER; MATTHES, 2024) Targeting the medical domain, `HealthFC` contains 750 bilingual (German and English) health claims labelled for veracity by medical experts, with evidence drawn from systematic reviews and clinical trials. Each claim is accompanied by annotated evidence sentences and a lay-language explanation paragraph for the verdict, supporting evidence retrieval, claim verification, and explanation generation.

Of these, we adopt `LIAR-RAW` for the evaluation in this chapter, retaining only the claims whose human-written justification can serve as a reference explanation. In Section 6.4 we detail the resulting subset and selection criteria.

6.3 Method

This section introduces the explainability extension of KG-CRAFT, augmenting it with a dedicated justification generation component built on the contrastive summary $A_{\mathcal{C}}$ produced by KG-CRAFT's contrastive reasoning pipeline (Section 4.3.3), and details the multi-desiderata evaluation protocol used to assess the quality of the resulting justifications. KG-CRAFT (Chapter 4) produces a veracity label for a claim from its evidence, whilst this chapter asks the verifier to additionally say *why*. We therefore frame the task as *verify-then-justify*, *i.e.*, the claim is first assigned a veracity label, and a natural language justification of that assignment is then produced from the same evidence (ATANASOVA et al., 2020; ZENG; GAO, 2024).

We begin by extending the problem statement from Section 4.3 to incorporate the justification generation task (Section 6.3.1). Subsequently, in Section 6.3.2, we describe how the explainability extension produces natural language justifications via a justification prompt p_{just}. Further, in Section 6.3.3, we introduce the justification quality assessment protocol, comprising the six desiderata and the LLM-as-a-Judge scoring procedure.

6.3.1 Problem Statement

Building on the claim verification formulation of Section 4.3, let $\mathcal{C}$ be a claim with an associated set of evidence reports $\mathcal{R}_\mathcal{C} = \{r_i\}_{i=1}^{|\mathcal{R}_\mathcal{C}|}$, and let $\mathcal{S}$ be a context derived from $\mathcal{R}_\mathcal{C}$ (instantiated as the distilled contrastive summary $A_\mathcal{C}$ for KG-CRAFT in Section 6.3.2, and as alternative summaries for the ablation conditions in Section 6.4). A verifier $f^V_{\theta_V} : (\mathcal{C}, \mathcal{S}) \mapsto \mathcal{V}_\mathcal{C}$ assigns the claim a veracity label $\mathcal{V}_\mathcal{C} \in \mathcal{Y}$ from this context. The objective of the explainable claim verification task is to additionally produce a natural language justification $\mathcal{J}_\mathcal{C}$ that justifies, in human-readable form, why the label $\mathcal{V}_\mathcal{C}$ was assigned to $\mathcal{C}$. We target a *justification* of the assigned label rather than a faithful account of the verifier's internal computation; we describe how this objective is met in Section 6.3.2.

We extend the verifier with a conditional justification map $f^J_{\theta_J} : (\mathcal{C}, \mathcal{V}_\mathcal{C}, \mathcal{S}) \mapsto \mathcal{J}_\mathcal{C}$ that generates the justification from the claim, the assigned label, and the same context $\mathcal{S}$ the verifier used. The explainable verifier is then

$$f_\theta(\mathcal{C}, \mathcal{R}_\mathcal{C}) = (\mathcal{V}_\mathcal{C}, \mathcal{J}_\mathcal{C}), \qquad \mathcal{V}_\mathcal{C} = f^V_{\theta_V}(\mathcal{C}, \mathcal{S}), \qquad \mathcal{J}_\mathcal{C} = f^J_{\theta_J}\big(\mathcal{C}, \mathcal{V}_\mathcal{C}, \mathcal{S}\big),$$

with parameters $\theta = (\theta_V, \theta_J)$ (instantiated as a single backbone model in Section 6.4). Conditioning f^J on $\mathcal{V}_\mathcal{C}$ encodes the verify-then-justify ordering, where the justification is generated after, and conditional on, the assigned label.

6.3.2 Justification Generation in KG-CRAFT

As the justification generation extension of KG-CRAFT, this section augments the framework introduced in Chapter 4 with a fourth component – *justification generation* – that builds on the contrastive summary $A_\mathcal{C}$ produced by the Answers Summarisation step (Section 4.3.3). KG-CRAFT's three original components (knowledge graph extraction, contrastive reasoning, and veracity verification) remain unchanged, as the verification step retains its label-only signature $\mathcal{V}_\mathcal{C} = LLM_{p_{\text{cv}}}(\mathcal{C}, A_\mathcal{C})$ as defined in Section 4.3.4. The justifier reuses the same backbone as the verifier, differing only in its prompt.

The justification is *post-hoc* and *label-conditioned*, *i.e.*, it is produced by a separate call, after and conditional on the assigned label.[1] We therefore present $\mathcal{J}_\mathcal{C}$ as a justification rather than a faithful account of the verifier's internal computation. Because the justifier is given the same contrastive summary $A_\mathcal{C}$ that the verifier used to assign the label, the justification is grounded in the evidence that determined the decision rather than reconstructed from the label alone.

We introduce a justification prompt p_{just} that takes the claim, the predicted veracity label, and the contrastive summary as inputs and produces a natural language justification:

$$\mathcal{J}_\mathcal{C} = LLM_{p_{\text{just}}}(\mathcal{C}, \mathcal{V}_\mathcal{C}, A_\mathcal{C}). \tag{6.1}$$

The justification is therefore semantically grounded in the contrastive summary $A_\mathcal{C}$, which itself distils the discriminating evidence between the claim and alternative entity substitutions extracted from the reports.

The justification step is intentionally decoupled from the classification step for modularity. As a consequence, the same p_{just} can be reused with alternative upstream summarisation strategies, allowing for a controlled study of how upstream design choices and provided context propagate into the quality of the resulting justification.

The proposed design in this section enables a direct ablation study of justification quality that parallels the veracity verification ablation of Section 4.6.1 (section 6.5.3). Furthermore, Section 6.5.3 empirically characterises the textual alignment between the generated justification, the contrastive summary, and the underlying evidence-marked report sentences.

6.3.3 Justification Quality Assessment

This section specifies how we assess the quality of the generated justifications. To this end, we first define a desideratum set $\mathcal{D}$. Subsequently, under the LLM-as-a-Judge procedure, we introduce a vector-valued quality function $q : \mathcal{J} \mapsto \{1, 2, 3, 4, 5\}^{|\mathcal{D}|}$, where each component $s_d(\mathcal{C})$ is the per-claim Likert score along desideratum $d \in \mathcal{D}$. The scores produced by this assessment are reported in Section 6.5.

Building upon the desiderata catalogue proposed by (KOTONYA; TONI, 2020b; WARREN; SHKLOVSKI; AUGENSTEIN, 2025) and on the broader XAI literature (GILPIN et al., 2018; DOSHI-VELEZ; KIM, 2017; MINH et al., 2022; SAEED; OMLIN, 2023), we instantiate $\mathcal{D}$ as

[1] Joint-decoding instantiations, in which the label and the justification are produced in a single pass (ATANASOVA et al., 2020), are out of scope for this chapter.

the following six dimensions, motivated and discussed in Section 6.2.1.

Definition 6.1 (Actionable)**.** *A justification is* actionable *when it provides factual corrections, identifies the specific elements of misinformation, references credible sources, and offers concrete contextual data sufficient for a downstream user to verify the assessment or take an informed decision.*

Definition 6.2 (Causal)**.** *A justification is* causal *when it establishes explicit causal chains and counterfactual reasoning between the available evidence and the assigned veracity label, articulating why the claim must be true or false rather than presenting purely correlational or descriptive information.*

Definition 6.3 (Contextful)**.** *A justification is* contextful *when it situates the claim in its surrounding evidence, provides self-contained reasoning that does not require external context to interpret, and integrates relevant background information without omission or fragmentation.*

Definition 6.4 (Impartial)**.** *A justification is* impartial *when it avoids partisan, ideological, or stylistic bias; relies on credible sources rather than rhetorical devices; and maintains internal logical consistency between the textual reasoning and the assigned veracity label.*

Definition 6.5 (Parsimonious)**.** *A justification is* parsimonious *when it conveys the discriminating evidence concisely, without redundant restatement of the claim, circular reasoning, or verbose filler that adds no evidentiary value.*

Definition 6.6 (Safety)**.** *A justification is* safe *when it does not reinforce, amplify, or restate the misinformation it purports to refute; nor does it introduce additional false information through hallucinated context or unsupported claims.*

These six desiderata are not independent. For instance, a hallucinated context may simultaneously degrade Impartiality (through internal inconsistency) and Safety (through introduction of unsupported information). The protocol nevertheless evaluates them separately to enable dimension-level diagnosis of method behaviour.

Each desideratum $d \in \mathcal{D}$ is operationalised through a dedicated evaluation prompt p_d that instructs an LLM judge to assess a candidate justification against the criteria of Definitions 6.1 to 6.6. The judge receives a four-tuple $(\mathcal{C}, \mathcal{V}_\mathcal{C}, E_\mathcal{C}, \mathcal{J}_\mathcal{C})$ comprising the claim, the assigned veracity label, the gold evidence sentences, and the candidate justification, respectively, and emits a Likert score $s_d(\mathcal{C}) \in \{1, 2, 3, 4, 5\}$ together with a textual reasoning field that articulates the basis for the assigned score.

The use of gold evidence sentences as the context provided to the judge aims to ensure that the judge evaluates the justification on the same evidential basis available to a human evaluator, rather than on raw reports, which may be lengthy, redundant, or partially irrelevant. The method-level score for a given (method, desideratum) pair is computed as the mean across the evaluated claims, $\bar{s}_d = \frac{1}{|\mathcal{C}|} \sum_{\mathcal{C}} s_d(\mathcal{C})$. To account for stochastic variation in LLM outputs, each evaluation is repeated across five independent runs, and the resulting scores are aggregated by mean and standard deviation.

In addition to method-level means, we measure the per-claim alignment between each candidate method and the reference explanation by computing Spearman's ρ, Pearson's r, and Kendall's τ correlation coefficients between the per-claim score vectors of the candidate method and those of the reference. Strong correlation indicates that the candidate method orders claims by explanation quality similarly to the reference, providing a complementary view to the method-level mean comparison.

6.4 Experimental Setup

We conduct a series of empirical evaluations to assess the quality of natural language justification for KG-CRAFT. In addition, we provide ablation studies that isolate the contribution of knowledge graph grounding and contrastive question formulation.

Dataset We use the `LIAR-RAW` (YANG, Z. et al., 2022) test split, restricted to the 855 claims with non-empty gold evidence (the original split contains 1,251 claims; claims without gold evidence cannot be scored against the reference explanation and are excluded). `LIAR-RAW` provides full-length PolitiFact ruling reports alongside human-written justification snippets, enabling a direct comparison between LLM-generated justifications and human-written counterparts on the same claim population. Among the datasets that pair fact-checked claims with annotated justifications (Section 6.2.4), `LIAR-RAW` is the largest, and its full-length ruling reports supply the rich evidence that KG-CRAFT depends on, since sparse reports yield sparser knowledge graphs and fewer contrastive questions, as identified in the report-length sensitivity of Section 4.8.

Compared Configurations We compare four LLM-based methods and the human-written reference. The LLM-based methods share the same Llama 3.3 70B backbone for both the verification and justification stages, instantiating the backbone setting described

in Section 6.3.1, so that observed differences reflect framework design rather than model capacity. Each method is described as:

- *Reference Explanation*: the human-written PolitiFact justification snippet associated with each claim. We use it as a reference rather than as an error-free standard, since the qualitative analysis of Section 6.5.4 shows it is itself sometimes fragmentary or non-explanatory;
- *Naïve LLM*: Llama 3.3 70B prompted with the claim and its reports, without extracted knowledge graph or contrastive reasoning;
- *Contrastive-LLM (no KG)*: the KG-CRAFT variant from Section 4.5.2.1 where contrastive questions are generated by the LLM through few-shot prompting rather than derived from the extracted knowledge graph;
- *KG + Reports (no contrastive)*: the KG-CRAFT variant from Section 4.6.1 where the LLM is augmented with the extracted knowledge graph but bypasses the contrastive question-answering step; and
- KG-CRAFT *(full)*: the complete framework with $\mathsf{K} = 5$ contrastive questions, exactly as evaluated in Section 4.5.1.

Intermediate Artefact Analysis In addition to the final justification produced by KG-CRAFT, we report reference-based metrics for the intermediate *Answers Summary* artefact (Section 4.3.3), enabling an analysis of how the justification prompt p_{just} (Section 6.3.2) reshapes the contrastive evidence into a final justification (Section 6.5.3).

Judge Model and Prompts We use Llama 3.3 70B as the judge for all six desiderata. We acknowledge that using the same model family for the verifier and the judge introduces a potential self-preference bias (CHEN, Z.-Y. et al., 2025; POMBAL; REI; MARTINS, 2026), but its effect is limited in our setting.[2] Because all four LLM-based methods share the same backbone whilst the reference explanation is human written, the bias is shared across

[2] We note that self-preference decreases sharply with model scale, for instance, Zhi-Yuan Chen et al. (2025) report a near-zero disentangled bias for Llama-3.1-70B (DBG $\approx 0.4\%$, versus 21.6% at 8B). Also, Pombal, Rei, and Martins (2026) report that the Llama family exhibits no self-overestimation (below-one HSPP ratio) on subjective rubrics. Thus, our judge model, Llama 3.3 70B, lies in the same large model, low-bias setting. Moreover, we score each justification on an absolute Likert scale against an explicit per-desideratum rubric rather than by pairwise preference, the paradigm in which self-preference is most severe (POMBAL; REI; MARTINS, 2026).

the compared methods, so the relative ranking remains valid as all methods are scored by the same judge under identical conditions. For the comparison against the reference it is moreover conservative, inflating the LLM methods relative to the reference so that the reported gap to human quality is, if anything, understated. Further discussion of this limitation appears in Section 6.7. The desideratum prompts (one per desideratum) are reproduced in Appendix A.2 and are designed to elicit a single Likert score together with a textual reasoning field.

Evaluation Metrics We report method-level Likert means and standard deviations per desideratum, aggregated over five independent judge runs. For per-claim alignment with the reference explanations, we report Spearman's ρ, Pearson's r, and Kendall's τ. We additionally report reference-based metrics (ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, and BERTScore F1) on the generated justification with respect to the reference explanation, enabling cross-checking between the LLM-as-a-Judge quality scoring and reference-based textual alignment.

6.5 Results and Analysis

This section is guided by the following research questions:

> **RQ1** How does KG-CRAFT compare to its ablations and to human-written reference explanations on the six proposed explanation desiderata?
>
> **RQ2** Does KG-CRAFT order claims by explanation quality similarly to the reference at the per-claim level?
>
> **RQ3** Do reference-based metrics corroborate the LLM-as-a-Judge assessment of justification quality?
>
> **RQ4** In the cases where the reference and KG-CRAFT disagree most strongly, do their lower-scored desiderata exhibit recurring, distinguishable failure patterns?

We address RQ1 in Section 6.5.1, RQ2 in Section 6.5.2, RQ3 in Section 6.5.3, and RQ4 in Section 6.5.4.

Table 10: LLM-as-a-Judge Likert scores (mean ± std over five judge runs) per desideratum and method on the `LIAR-RAW` test set ($n = 855$). All methods use Llama 3.3 70B as the verifier and as the judge. Best non-reference scores per desideratum in **bold**.

Desideratum	Reference Explanation	Naïve LLM	Contrastive-LLM (no KG)	KG + Reports (no contrastive)	KG-CRAFT (full)
Actionable	2.960 ± 0.003	2.323 ± 0.006	2.423 ± 0.006	2.349 ± 0.008	**2.478 ± 0.007**
Causal	2.220 ± 0.005	1.757 ± 0.003	**1.855 ± 0.007**	1.796 ± 0.007	1.806 ± 0.005
Contextful	3.250 ± 0.006	2.740 ± 0.005	2.850 ± 0.002	2.716 ± 0.007	**2.884 ± 0.002**
Impartial	3.429 ± 0.009	3.025 ± 0.001	3.022 ± 0.001	3.023 ± 0.001	**3.221 ± 0.012**
Parsimonious	2.620 ± 0.008	2.024 ± 0.003	2.053 ± 0.008	2.054 ± 0.003	**2.143 ± 0.003**
Safety	3.976 ± 0.009	3.691 ± 0.004	3.743 ± 0.008	3.668 ± 0.006	**3.796 ± 0.009**

6.5.1 Justification Quality across Desiderata

We first address RQ1, presenting the method-level LLM-as-a-Judge results in Table 10. Across all six desiderata, the human-written reference explanation achieves the highest absolute Likert mean, indicating the protocol's construct validity. The smallest absolute gap between the reference and the best non-reference method is on *Safety* (3.976 ± 0.009 vs. 3.796 ± 0.009; −0.18 points), suggesting that cautious, non-amplifying language is a property of the shared backbone, common to all methods built on it. The largest absolute gap is on *Causal* (2.220 ± 0.005 vs. 1.855 ± 0.007; −0.36 points), reproducing the long-standing finding that abstractive causal language is the hardest aspect of explanation generation (GILPIN et al., 2018).

Overall analysis shows KG-CRAFT winning five of the six desiderata against every ablation: Actionable (+0.10 points over the next-best ablation), Contextful (+0.04), Impartial (+0.20), Parsimonious (+0.09), and Safety (+0.05). The Impartial gap is the most pronounced, where all three non-KG-CRAFT methods sit at approximately 3.02, whilst full KG-CRAFT jumps to 3.221. This pattern indicates that grounding the reasoning in extracted knowledge graph triples reduces the verifier's exposure to biases or stylistic phrasing in the raw reports, an effect not achieved by either knowledge graph augmentation alone or contrastive reasoning alone.

The exception to this pattern is the *Causal* dimension, where the Contrastive-LLM-no-KG variant outperforms KG-CRAFT (1.855 ± 0.007 vs. 1.806 ± 0.005; −0.05). We interpret this as a discovered dimension-specific trade-off rather than a methodological flaw. The unconstrained nature of LLM-generated contrastive questions produces a wider hypothesis space, which the LLM then converts into more explicitly causal phrasing. KG-grounded contrastive substitutions, whilst more faithful to the available evidence, may be more declarative than causal in tone.

Expressed as a percentage of the reference, KG-CRAFT achieves 83.7% on Actionable, 81.3% on Causal, 88.8% on Contextful, 93.9% on Impartial, 81.8% on Parsimonious, and 95.5% on Safety (an arithmetic mean of approximately 87.5% across the six desiderata), demonstrating that KG-CRAFT's natural language justifications are closer to the human-written reference quality whilst additionally delivering the substantial veracity-assessment improvements documented in Section 4.5.1.

A noteworthy contrast with Chapter 4 concerns the relative ordering of the ablations. For verification F1, Table 5 establishes the ordering KG-CRAFT $\gg$ KG+Reports $>$ Contrastive-LLM-no-KG $\sim$ Naïve. For explanation Likert means, the ordering is instead KG-CRAFT $>$ Contrastive-LLM-no-KG $>$ KG+Reports $\sim$ Naïve. We interpret this dissociation as evidence for a component-attribution argument, which *contrastive question formulation* is the component that primarily drives explanation quality, whilst *knowledge graph grounding* is the component that primarily drives verification accuracy. The full framework benefits from both, achieving simultaneous improvements on accuracy and on explanation quality without requiring a trade-off between the two.

6.5.2 Per-claim Alignment with Reference Explanations (Correlations)

Whilst method-level Likert means establish that KG-CRAFT produces higher-quality justifications on average, they do not reveal whether the method orders individual claims by explanation quality in the same way as the reference. RQ2 addresses this question through correlation analysis between per-claim score vectors.

Table 11 reports Spearman's ρ, Pearson's r, and Kendall's τ between each candidate method and the reference, per desideratum. All values are statistically significant at $p \leq 0.05$ unless otherwise indicated. The three coefficients capture complementary notions of agreement. Kendall's τ measures the probability that any two claims are ordered the same way by a candidate method and by the reference, minus the probability that they are ordered oppositely; since RQ2 concerns whether KG-CRAFT *orders* claims by explanation quality as the reference does, τ is the most directly aligned measure, and it is robust to small perturbations in the scores. We additionally report Spearman's ρ, which measures monotonic rank agreement, and Pearson's r, which measures linear association between the raw score magnitudes; ρ corroborates τ on the ordering, whilst r imposes the stricter requirement that the magnitudes also align. We therefore lead the analysis with τ and use ρ and r as corroboration.

Two patterns emerge. First, *Contextful* is consistently the most reproducible desider-

Table 11: Per-claim alignment with the reference explanation: Spearman's ρ, Pearson's r, and Kendall's τ between each candidate method's per-claim Likert score vector and the reference's vector, on the `LIAR-RAW` test set. All values significant at $p \leq 0.05$. Best per (desideratum, coefficient) in **bold**.

Method	Action.	Causal	Context.	Impart.	Parsim.	Safety
Spearman's ρ						
Naïve LLM	0.256	0.191	0.409	**0.221**	**0.138**	0.212
Contrastive-LLM (no KG)	**0.333**	**0.258**	0.418	0.211	0.129	**0.226**
KG + Reports (no contrastive)	0.254	0.199	0.399	0.175	0.108	0.103
KG-CRAFT (full)	0.310	0.224	**0.428**	0.177	0.113	0.218
Pearson's r						
Naïve LLM	0.268	0.174	0.408	**0.219**	**0.139**	0.215
Contrastive-LLM (no KG)	**0.331**	**0.261**	0.410	0.213	0.121	**0.225**
KG + Reports (no contrastive)	0.253	0.219	0.391	0.172	0.108	0.099
KG-CRAFT (full)	0.304	0.235	**0.422**	0.179	0.121	0.202
Kendall's τ						
Naïve LLM	0.222	0.176	0.361	**0.193**	**0.126**	0.201
Contrastive-LLM (no KG)	**0.289**	**0.240**	0.368	0.186	0.118	**0.215**
KG + Reports (no contrastive)	0.222	0.183	0.351	0.152	0.099	0.098
KG-CRAFT (full)	0.270	0.206	**0.378**	0.155	0.102	0.207

atum at the claim level, with Kendall's τ ranging from 0.351 to 0.378 across the four ablations (Spearman's ρ from 0.399 to 0.428, Pearson's r from 0.391 to 0.422), and KG-CRAFT attaining the highest value ($\tau = 0.378$, $\rho = 0.428$, $r = 0.422$). This is the only desideratum where the judge agrees with the reference explanation on *which claims* deserve a richer contextualisation. Second, *Parsimonious* is consistently the least reproducible, with $\tau \leq 0.126$ ($\rho \leq 0.138$, $r \leq 0.139$) across all methods. Combined with the low absolute Likert means observed in Table 10, this suggests that parsimony is intrinsically difficult both to produce and to score reliably for fact-checking justifications. Kendall's τ is uniformly lower than Spearman's ρ throughout the table, as expected from its pairwise definition, so the two are not directly comparable in magnitude. On the question that RQ2 poses, all three coefficients largely converge, identifying the same best-aligned method in every desideratum, and the two rank-based measures (τ and ρ) agree on the full ordering of methods throughout, whilst the only difference between the rank-based and magnitude-based measures is a lower-rank reshuffle among the non-leading methods on *Safety*.

For the *Causal* dimension, the per-claim correlations corroborate the method-level finding, where Contrastive-LLM (no KG) shows the highest agreement ($\tau = 0.240$, $\rho = 0.258$, $r = 0.261$) whilst KG-CRAFT shows $\tau = 0.206$ ($\rho = 0.224$, $r = 0.235$), reinforcing the dimension-specific trade-off identified in Section 6.5.1.

The self-correlation of the reference against itself is $\rho = r = \tau = 1.0$ across all desiderata (not reported in Table 11), confirming the consistency of the five-run aggregation procedure.

6.5.3 Textual Alignment with Reference-based Metrics

In RQ3, we exam whether reference-based metrics corroborate the LLM-as-a-Judge assessment of justification quality. We approach this by measuring how closely each text overlaps with the evidence-marked sentences, the shared source from which both the human-written reference explanation and the LLM justifications are derived. Our findings show the lexical-overlap metrics (ROUGE, BLEU) contradict the judge, *i.e.*, the human-written reference explanation overlaps the evidence far less than the KG-CRAFT justification does, even though the judge rates the reference highest. The semantic metric (BERTScore) brings the two close (0.279 vs. 0.288), but still does not reproduce the judge's clear preference for the reference. Table 12 reports ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, and BERTScore F1 for the human-written reference explanation, the KG-CRAFT Generated Justification, and the intermediate Answers Summary, each measured against the evidence-marked sentences.

Table 12: Reference-based metrics on the `LIAR-RAW` test set, comparing the reference explanation, the KG-CRAFT (Llama 3.3 70B) Generated Justification, and the intermediate Answers Summary against the dataset's evidence-marked report sentences (*i.e.*, the subset of report sentences flagged as gold evidence by `LIAR-RAW`). R1, R2, RL: ROUGE-1, ROUGE-2, ROUGE-L (averaged variant). B4: BLEU-4. BS-F1: BERTScore F1. The first row reports the reference explanation's distance from the evidence-marked sentences, serving as a human-abstraction baseline.

Subject	**R1**	**R2**	**RL**	**B4**	**BS-F1**
Reference Explanation (vs. Evidence)	0.082	0.025	0.058	0.237	0.279
KG-CRAFT Generated Justification	0.180	0.036	0.125	0.580	0.288
KG-CRAFT Answers Summary (intermediate)	0.170	0.038	0.115	0.526	0.179

The lexical overlap of the KG-CRAFT Generated Justification with the underlying reports is approximately twice that of the reference explanation on the lexical metrics (ROUGE-1: 0.180 vs. 0.082; ROUGE-L: 0.125 vs. 0.058; BLEU-4: 0.580 vs. 0.237). By contrast, the semantic metric BERTScore F1 is near-identical for the two (0.288 vs. 0.279), so this gap is specific to lexical overlap. This asymmetry is itself corroborated by the construction of the `LIAR-RAW` reference explanations, which are characteristically paraphrastic, summarising, and recontextualising rewrites of the underlying PolitiFact reports rather than direct quotations from them (their low ROUGE/BLEU values against

the evidence-marked sentences are therefore a property of the dataset, not an artefact of the metrics). The pattern shows that LLM-generated justifications retain greater textual alignment with the input evidence, whilst the human-written reference explanations diverge more from it. This divergence indicates that the reference explanations are more *abstractive*, restating the evidence in their own vocabulary rather than reusing the evidences' vocabulary. More importantly, high lexical overlap with reports is therefore *not* a marker of explanation quality on this task. Therefore, using ROUGE or BLEU as the sole evaluation criterion would systematically reward extractive over abstractive text, contradicting the judge's assessment of which justifications are better. This finding validates the methodological choice of LLM-as-a-Judge for explanation evaluation in fact-checking and is consistent with the parallel observation in Section 4.5.2.1 that KG-based grounding boosts evidence-aligned scores (AlignScore, RQUGE) without proportionally boosting reference-based textual alignment metrics.

Second, comparing the intermediate Answers Summary against the final Generated Justification (BLEU-4: 0.526 $\rightarrow$ 0.580; ROUGE-L: 0.115 $\rightarrow$ 0.125; BERTScore F1: 0.179 $\rightarrow$ 0.288), the justification prompt p_{just} (Section 6.3.2) produces text that retains slightly more of the evidence-marked sentences' wording than the contrastive summary it consumes. The Answers Summary captures the contrastive content at a higher abstractive register, and the justification step recovers a more extractive register (*i.e.*, more report-aligned) whilst preserving the contrastive content. Even with this gain, the textual alignment recovered by p_{just} is with the evidence-marked sentences, not with the reference explanation. The reference remains more abstractive at $R_1 = 0.082$, reinforcing the findings of Section 6.5.1 that human fact-checking is markedly more abstractive than every LLM-generated method.

6.5.4 Qualitative Failure Pattern Analysis

The quantitative results presented in Section 6.5.1 establish that KG-CRAFT produces higher-quality justifications than its ablations on five of six desiderata. However, these results do not explain *why* the win is uneven across desiderata or where the single loss on *Causal* originates. RQ4 addresses this through a structured qualitative analysis of 19 disagreement cases, defined as claims in which the reference and KG-CRAFT Likert scores differ by at least 2 points on at least one desideratum. For each case, we examine the candidate justifications produced by all five methods together with the textual reasoning fields generated by the judge.

Table 13 reports six exemplars drawn from the 19 cases (one per recurring failure pattern) together with each method's Likert score on the relevant desideratum. The remaining 13 cases are referenced inline in the failure-pattern discussion that follows. The CL, Naive, and KGR columns report bystander scores from the three ablation methods that were not used to define the disagreement set. On reference-bad cases (rows 1-3), the bystanders cluster closer to the reference than to KG-CRAFT, indicating that KG-CRAFT's superior score on these cases is not a judge artefact shared by other LLM-generated justifications. On KG-CRAFT-bad cases (rows 4-6), the bystanders cluster closer to KG-CRAFT than to the reference, suggesting that the identified failure modes are largely shared across LLM-generated methods rather than introduced by KG grounding.

Table 13: Six exemplars from the 19 disagreement cases (one per recurring failure mode) on the `LIAR-RAW` test set. For each case, the column under each method reports its Likert score on the relevant desideratum.

ID	Claim (truncated)	Desid.	Ref.	KG-CRAFT	CL	Naive	KGR	Failure mode
3399	"EPA ruling forces dairy farmers to comply with Spill Prevention regulations..."	Action.	1	3	2	2	2	Ref.: rhetorical but non explanatory
2470	"...a security clearance even the president cannot obtain because of my background."	Causal	1	4	2	3	2	Ref.: single-sentence assertion
4243	"The U.S. has the longest surviving constitution."	Parsim.	2	4	2	2	2	Ref.: redundant journalistic prose
5249	"Gov. McDonnell's proposed budget is cutting public education."	Impart.	4	1	2	2	2	KG-CRAFT: internal logical contradiction
769	"Barack Obama ...96 percent of his votes have been solely along party line."	Context.	4	1	2	1	2	KG-CRAFT: hallucinated context
6604	"When Tim Kaine was governor, spending soared, blowing holes in the budget every year."	Causal	4	1	3	2	2	KG-CRAFT: circular restatement

In the following, we discuss three failure patterns observed in the reference explanations, three characteristics of KG-CRAFT justifications, and conclude with an observation about the relationship between the two.

Reference Explanation Failure Patterns First, reference explanations sometimes adopt a rhetorical but non-explanatory pattern that the judge penalises. For instance, the reference explanation for Claim 3399 (Table 14) reads "*Griffith is dishing udder cow chips*", which provides no factual correction. The judge's reasoning explicitly catalogues five missing components: factual corrections, specific misinformation elements, credible sources, contextual data and regulatory details, and substantive engagement with the claim. Second, reference explanations are sometimes disconnected snippets extracted from longer journalistic articles. For Claim 2135 (Schilling bloody sock; Table 14), the reference explanation is a Chafee-campaign joke ("*We shot ourselves in the sock*") that does not address the claim, *i.e.*, whether a teammate said the sock was painted. The judge notes that the explanation "*appears disconnected from the actual claim being fact-checked*". A third observed failure pattern in reference explanations consists of single-sentence assertions without a causal chain. For Claim 2470 (Table 14), the reference explanation is one sentence "*The president has the highest security clearance in the land*", which is true but neither counterfactual nor causal. Collectively, these patterns reflect that PolitiFact reports are written for human readers and structured as journalistic narratives, and the reference explanations in `LIAR-RAW` are short summarising snippets of this narrative rather than stand-alone explanations. The gap between the reference and KG-CRAFT on these cases demonstrates a meaningful difference between the generated justifications and journalistic prose.

KG-CRAFT Failure Patterns The first observed failure pattern on KG-CRAFT generated justifications exhibits an internal logical contradiction. For instance, in Claim 5249 (a FALSE claim about budget cuts; Table 14), the KG-CRAFT justification reads "*The claim has been classified as FALSE because the budget actually includes significant cuts to public education, totaling $646 million*", *i.e.*, the textual justification supports the opposite of the assigned label. The judge identifies this contradiction explicitly and assigns an Impartial score of 1 (the reference scores 4 on the same dimension). Second, KG-CRAFT sometimes hallucinates context that is not present in the evidence. Claim 769 (Table 14) exposes this failure pattern, where KG-CRAFT remarks the assertion "*Obama receiving 96% of votes from black women*", which is not part of the provided input reports. The judge enumerates five concrete factual fabrications and assigns Contextful and Impartial scores of 1 each (the reference scores 4 on both). Third, KG-CRAFT sometimes produces circular restatement of the assigned label without providing evidence-grounded reasoning,

as exemplified by Claim 6604 (Table 14) where the justification reformulates the "*BARELY TRUE*" label without engaging with the spending versus revenue distinction central to the gold reasoning.

After analysing the failure patterns of reference explanations and KG-CRAFT, we observed they are largely disjoint. That is, the structural failure patterns found on reference explanations are rhetorical, fragmentary, and single sentence, which concentrate on dimensions that reward self-contained, structured presentation (Actionable, Parsimonious, and Contextful); whilst the structural failure patterns found on KG-CRAFT justifications are contradiction, hallucination, circularity, which concentrate on dimensions that demand either extended counterfactual reasoning chains (Causal) or strict separation between evidence-text and claim-text (Impartial). The asymmetry failure patterns across the desiderata corroborates to the findings of Section 6.5.1. Furthermore, due to their non-overlapping characteristics, it suggests that the two paradigms (human-written and KG-CRAFT-generated justifications) could be combined fruitfully rather than substituted for one another. We additionally observe a veracity label asymmetry in the disagreement cases, where the pair bad reference explanations and good KG-CRAFT justifications cases skew towards decisive labels (FALSE, PANTS FIRE, BARELY TRUE), where KG-CRAFT's structured contrastive reasoning has a clear verdict to defend. Conversely, the pair of good reference explanations and bad KG-CRAFT justifications cases skew towards graded labels (MOSTLY TRUE, HALF TRUE), where the justification must articulate why the claim is partially true, requiring engagement with both sides of the qualification, an inductive bias not directly served by contrastive entity substitution.

We caution that the patterns aforementioned are extracted from a curated subset of 19 extreme-disagreement cases. The patterns characterise the occurrence of failure patterns rather than their prevalence, *i.e.*, these patterns should be understood as they recur in low-scoring cases rather than as the primary causes of KG-CRAFT's failures. The judge's reasoning fields, which we quote verbatim throughout this subsection, are themselves LLM output and inherit the same self-preference bias as the Likert scoring.

6.6 Conclusion and Discussions

In this chapter, we proposed a multi-desiderata evaluation protocol for LLM-generated fact-checking justifications, operationalising six standard XAI desiderata via LLM-as-a-Judge on a Likert scale, and applied it to KG-CRAFT alongside a four-method ablation suite on

the LIAR-RAW test set ($n = 855$). Our experiments demonstrate that KG-CRAFT's natural language justifications outperform every ablation on five of six desiderata (Actionable, Contextful, Impartial, Parsimonious, and Safety) and reach approximately 87.5% of the human-written reference explanation quality on average. The largest improvement is observed on the *Impartial* dimension (3.221 ± 0.012 for KG-CRAFT versus approximately 3.02 for all three ablations), evidencing that knowledge graph grounding measurably reduces the verifier's exposure to biased or stylistic phrasing of raw reports. The single exception is the *Causal* dimension, where the Contrastive-LLM-no-KG ablation marginally outperforms KG-CRAFT (1.855 versus 1.806). A key methodological observation is the dissociation between the veracity verification performance ranking documented in Chapter 4 (KG-CRAFT $\gg$ KG+Reports $>$ Contrastive-LLM-no-KG $\sim$ Naïve) and the explanation quality ranking documented in this chapter (KG-CRAFT $>$ Contrastive-LLM-no-KG $>$ KG+Reports $\sim$ Naïve), which is indicative that contrastive question formulation primarily drives explanation quality, whilst knowledge graph grounding primarily drives verification accuracy, and the full framework benefits from both simultaneously. These findings directly address Research Question 4, showing that, the KG-grounded contrastive configuration approaches the quality of human-written reference explanations on aggregate across the six desiderata. At the per-claim level, Contextfulness emerges as the most reproducible desideratum (Spearman's $\rho \approx 0.43$ with the reference) and Parsimony as the least ($\rho \leq 0.14$), informing future LLM-as-a-Judge protocol design. Comparing the LLM-as-a-Judge scores against reference-based metrics, we find that lexical-overlap metrics (ROUGE, BLEU) contradict the judge's quality assessment, rating the abstractive reference explanations below the more extractive LLM justifications, which validates the choice of LLM-as-a-Judge over reference-based evaluation for this task. Finally, our structured failure pattern analysis on 19 disagreement cases reveals that human reference explanations and KG-CRAFT justifications fail in largely non-overlapping failure patterns. Whilst reference explanations are sometimes rhetorical, fragmentary, or non-causal snippets, KG-CRAFT justifications occasionally exhibit internal logical contradiction, hallucinated context, or circular restatement.

There are several promising directions for future work to build upon these findings. First, we aim to validate the LLM-as-a-Judge protocol against human assessment, extending the small-scale human evaluations conducted in prior work (KOTONYA; TONI, 2020a) to the multi-desiderata setting introduced here, and quantifying the residual self-preference bias from using a backbone-family judge. Second, multi-judge protocols leveraging panels of LLMs from different model families would directly mitigate self-preference bias and

provide a more robust evaluation. We hypothesise that the relative ranking of methods would remain stable but the absolute Likert magnitudes would shift, particularly on the most subjective desiderata. Third, the discovered Causal trade-off motivates a targeted prompt-engineering step that explicitly elicits causal language during the verification prompt. We hypothesise that this could close the Causal gap without sacrificing the wins on the other five dimensions. Fourth, the qualitative failure patterns analysis suggests that KG-CRAFT would benefit from an automatic self-consistency check at generation time, where the verifier is re-prompted whenever its justification text disagrees with its assigned label (addressing Claim 5249-style Table 14 contradictions). Furthermore, evidence-attribution prompting to mitigate hallucinations (addressing Claim 769-style Table 14 fabrications). We believe these two mechanism would enable KG-CRAFT as a solution in unconstrained real fact-checking environment. Fifth, extending the evaluation to additional fact-checking benchmarks beyond `LIAR-RAW` (particularly those with available reference explanations across different domains and languages) would strengthen the generalisability of the findings. Finally, the application of the proposed six-desideratum protocol to other XAI evaluation contexts beyond automated fact-checking, such as medical decision support or legal-reasoning systems, would validate the protocol's broader utility and contribute to the wider XAI evaluation literature.

6.7 Limitations

The findings reported in this chapter are subject to two categories of limitations, those inherited from the LLM-as-a-Judge evaluation paradigm and those specific to the design and scope of the present study.

When the judge model and the verifier model belong to the same family, the judge may systematically inflate scores for outputs that are stylistically similar to its own generation, distorting the absolute Likert magnitudes, this limitation is called *self-preference bias*. The relative ordering of methods remains valid when all methods are scored by the same judge under identical conditions, but the absolute "percentage of the reference" figure should be interpreted accordingly. A second concern is *prompt sensitivity*. Small changes in the wording of the desideratum prompts can shift the rank ordering of compared methods, and multi-prompt ensembling in which several semantically equivalent variants of each prompt are aggregated would mitigate this sensitivity at additional computational cost. The 5-point Likert scale used here is also a source of *measurement coarseness*, since it cannot capture fine-grained quality differences and tends to compress scores towards the centre.

Higher resolution scales, ranking-based protocols, or pairwise comparison may reveal additional structure not visible in the means reported here. Finally, the textual reasoning fields generated by the judge, on which our qualitative analysis draws, are themselves *produced by a large language model* and inherit the same biases as the Likert scoring. They should be read as judge-articulated rationales rather than as external evidence.

Other limitations are worth noting, organised here into three groups of related constraints. The first concerns the *scope of the evaluation*. We assess explanation quality only on the `LIAR-RAW` test set, restricted to claims with non-empty gold evidence, and run the headline ablation with Llama 3.3 70B as both verifier and judge. The dataset choice is forced by data availability, since comparable human-written reference explanations are limited in other datasets, and a single-backbone setup keeps the LLM-as-a-Judge inference budget tractable across the four ablations and five repeated runs. The trade-off is that absolute magnitudes inherit some self-preference risk and that domain and language generalisation remain open, both of which are identified as future work in Section 6.6. The second concerns *bias mitigation*. Five independent judge runs are sufficient to suppress stochastic noise but not systematic bias, and LLM-as-a-Judge is used as a scalable proxy rather than as a replacement for *human evaluation*, which remains the gold standard for explanation quality. A multi-judge panel comprising different model families would directly address the systematic-bias dimension, and a calibrated human-evaluation study would anchor the absolute Likert scores reported here. The choice to defer both to future work reflects the cost of multi-judge inference and the impracticality of recruiting domain-expert annotators at the scale of the present evaluation. The third concerns the *depth of the qualitative failure pattern analysis*. Section 6.5.4 draws on 19 deliberately curated extreme-disagreement cases, and the recurring patterns it surfaces (logical contradiction, hallucinated context, circular restatement) are identified through manual inspection of the judge's reasoning fields. Curation gives us mechanism-level insight but leaves the prevalence of each pattern not measured, limiting automated detection at scale. Both extensions, namely population-level prevalence estimation and deployment time consistency-checking pipelines, are identified as future work in Section 6.6.

7 Conclusion

This thesis investigated how automated fact-checking systems can leverage multimodal integration and structured reasoning to improve both veracity assessment and the quality of the explanations that accompany it. The work was organised around two framework families, each paired with an explainability extension. In Chapters 3 and 5, we propose mu2 and its explainable counterpart mu2X, respectively, addressing classification-based misinformation detection over multimodal and multilingual social network content. In Chapters 4 and 6, we propose KG-CRAFT and its justification generation extension, respectively, addressing evidence-based claim verification through LLM-driven knowledge graph-based contrastive reasoning.

Throughout, the central premise was that the diversity of evidence that drives accurate detection, whether across modalities for classification or across structured knowledge relations for verification, is the same diversity that supports richer and more interpretable explanations. The two challenges identified in Chapter 1, namely the integration of heterogeneous and structured information and the explainability of the resulting decisions, were therefore addressed jointly rather than in isolation. Addressing the research questions outlined in Chapter 1, we hereby present the conclusions derived from this thesis.

Multimodal Integration for Misinformation Detection This thesis establishes that integrating textual content with social network relational structure improves classification-based misinformation detection across diverse linguistic settings (Research Question 1). Through mu2, which couples language-specific text encoders with graph attention networks over post-node representations, multimodal fusion achieves the best classification performance in every evaluated language of the `MuMiN` benchmark, with a multilingual F1-score of 0.9738 and per language scores reaching 0.9951 on Portuguese. A cross-lingual analysis further reveals that textual features are more robust than graph-based representations in low-resource and sparse graph scenarios, whilst the two modalities perform comparably in isolation. This has practical implications for multilingual deployment: text-based encoders

provide a reliable baseline, while multimodal fusion delivers consistent gains wherever relational data is available.

Knowledge Graph Contrastive Reasoning for Claim Verification This thesis demonstrates that grounding LLM-based claim verification in knowledge graphs extracted from claim-associated reports and reasoning over them through contrastive questions, substantially improves veracity assessment over direct prompting and unstructured retrieval-augmented baselines (Research Question 2). KG-CRAFT achieves state-of-the-art F1-scores of 73.87% on `LIAR-RAW` and 81.58% on `RAWFC`, and our ablations isolate knowledge graph grounding as the primary driver of this performance. Ablations that replace the KG-grounded contrastive questions with unstructured LLM-generated alternatives collapse performance from 73.87% to 29.68% F1 on `LIAR-RAW`. The framework generalises beyond political fact-checking, reaching 83.03 F1 on `SciFact` and 78.66 balanced accuracy on `PubHealth`, and its contrastive evidence summaries serve as an effective form of knowledge distillation, enabling a 1.7B-parameter SmolLM2 model to reach 73.40% F1 despite being orders of magnitude smaller than the naïve LLM backbones it outperforms.

Trustworthiness and Robustness of Multimodal Explanations This thesis shows that multimodal feature integration is itself a source of explanation quality, not only of predictive accuracy (Research Question 3). Extending mu2 into mu2X with a Classification Explainability Module that combines GraphLIME for relational features and Integrated Gradients for textual features, and introducing two evaluation protocols beyond standard interpretability, we find that multimodal explanations are the most stable across top-K feature selections, achieving the best F1-scores at both top-1 (0.8207) and top-10 (0.8115). Results also show a complementary profile, where text-based explanations are the most resilient to the injection of noisy features, graph-based explanations the least, and multimodal explanations occupy an intermediate position that softens the non-linear behaviour of the relational signal. The interpretability analysis additionally identifies engagement metadata as the most frequent single feature explanation across all languages, with low-resource languages relying more heavily on metadata and graph-based features owing to sparser relational neighbourhoods.

Quality of Knowledge-Grounded Natural Language Justifications This thesis establishes that the structured contrastive reasoning supporting KG-CRAFT also provides

the foundation for high-quality natural language justifications (Research Question 4). Applying six standard explainability desiderata through an LLM-as-a-Judge protocol on the `LIAR-RAW` test set, we find that KG-CRAFT's justifications outperform every ablation on five of the six desiderata and reach approximately 87.5% of the human-written reference explanation quality on average, with the largest gain on the *Impartial* dimension, evidencing that knowledge graph grounding measurably reduces exposure to biased or stylistic phrasing of raw reports. The single exception is the *Causal* dimension, where an ablation without knowledge graph grounding marginally wins. A key methodological observation is the dissociation between the verification accuracy ranking of Chapter 4 and the explanation quality ranking of Chapter 6, where contrastive question formulation primarily drives explanation quality, whilst knowledge graph grounding primarily drives verification accuracy, and the full framework benefits from both simultaneously.

Revisiting the Hypotheses Taken together, these findings support the main hypothesis of this thesis: across both framework families, integrating heterogeneous evidence and reasoning over its relational structure enables systems that are simultaneously more accurate and more explainable. Clause (i) is supported by mu2 and mu2X, where multimodal fusion improves cross-lingual detection and yields the most stable feature-level explanations, with each modality contributing a complementary robustness property. Clause (ii) is validated by KG-CRAFT and its justification generation extension, where knowledge graph contrastive reasoning improves verification over direct prompting and retrieval-augmented baselines, and the same contrastive structure produces justifications that approach the human-written reference quality. The dissociation between the drivers of accuracy and of explanation quality is a nuance the original hypothesis did not anticipate, and it reinforces the thesis's broader argument that predictive performance and explanation quality, whilst correlated, are addressed by partly distinct design choices and are best pursued jointly.

7.1 Future Work

Having summarised the key findings and building upon them, we outline promising directions for future work, which apply across the contributions of this thesis.

Human-Centred and Multi-Judge Explanation Evaluation The explanation quality

results of this thesis rest on simulated and automated proxies, *i.e.*, a linear surrogate user for the trustworthiness protocol of Chapter 5, and a single-family LLM-as-a-Judge for the desiderata protocol of Chapter 6. Both are scalable but inherit known biases, namely the gap between simulated and real user behaviour and the self-preference of same-family judges. A natural next step is to anchor these protocols against calibrated human evaluation and to replace the single judge with a panel of LLMs drawn from different model families. We expect the relative ranking of methods to remain stable whilst absolute magnitudes shift, particularly on the most subjective desiderata, and such studies would establish how faithfully the automated protocols approximate human judgements of explanation quality.

Unified Feature-Level and Natural Language Explanations This thesis advances two complementary explanation paradigms, feature-level attributions in Chapter 5 and knowledge grounded natural language justifications in Chapter 6, but treats them separately. A promising direction is to integrate the two, grounding human-readable justifications in the feature attributions that drive a classification decision, so that an explanation states not only *why* a verdict was reached but also *which* signals were most responsible. Such integrated explanations would move beyond feature-level scores, which demand technical literacy to interpret, towards justifications that are simultaneously faithful to the model and actionable for content moderators and end users.

Robust Deployment over Constrained and Evolving Evidence The frameworks proposed here are evaluated using static benchmarks with relatively rich evidence, yet several findings point to challenges in real-world deployment. mu2 is trained on fixed network snapshots and does not exercise the temporal dynamics of evolving misinformation, whilst KG-CRAFT exhibits a report-length sensitivity, where short evidence yields sparser knowledge graphs and a smaller contrastive question space. Future work should investigate adaptive approaches that track shifts in content and network topology over time, and verification pipelines that remain effective over short-form claims where evidence is constrained at retrieval time. The qualitative failure patterns of Chapter 6 further motivate generation-time safeguards, such as self-consistency checks that re-prompt the verifier whenever its justification disagrees with its assigned label and evidence-attribution prompting to mitigate hallucinated context, both prerequisites for deployment in unconstrained fact-checking environments.

Cross-Lingual and Cross-Domain Generalisation Although this thesis spans three languages and four domains, broader generalisation remains open across all four contributions. The classification frameworks are confined to the three languages of `MuMiN`, the only widely adopted multilingual benchmark combining the required social network modalities, and the LLM-based frameworks are evaluated primarily on English data, with intermediate components such as entity linking and relation extraction liable to language-specific failure modes. Extending the evaluation to typologically diverse and lower-resource languages, and to benchmarks with different annotation schemes, network structures, and evidence characteristics, would strengthen the generalisation capabilities of both the frameworks and the proposed evaluation protocols. The application of the trustworthiness, robustness, and desiderata protocols to explainable systems beyond misinformation detection, such as medical decision support or legal reasoning, would likewise validate their broader utility within the XAI evaluation literature.

Foundation Models and Efficient Deployment Finally, the architectural choices of this thesis invite revisiting due to the rapidly advancing foundation models. Replacing mu2's language-specific encoders with multilingual foundation models could improve representation quality and extend language coverage without dedicated per language pre-training, whilst KG-CRAFT's demonstrated ability to distil information through contrastive reasoning applied to small language models motivates lightweight deployment configurations, including quantised and on-device inference, to broaden accessibility for practitioners without large-scale inference infrastructure.

7.2 Summary

This thesis set out to make automated fact-checking systems both more accurate and more explainable by leveraging multimodal integration and structured reasoning, and it pursued this goal across two framework families spanning classification-based detection and evidence-based verification, each paired with an explainability extension. The contributions show that multimodal integration improves cross-lingual detection and yields the most stable feature-level explanations; that knowledge graph contrastive reasoning delivers state-of-the-art verification and supports justifications approaching human-written quality; and that predictive performance and explanation quality, whilst connected, are driven by partly distinct design choices and are best pursued together. We close with a vision for

the directions outlined above, namely human-centred and multi-judge evaluation, unified feature-level and natural language explanations, robust deployment over constrained and evolving evidence, cross-lingual and cross-domain generalisation, and the adoption of foundation models for efficient deployment, in the hope that they bring automated fact-checking closer to the transparency and reliability that its real-world stakes demand.

REFERENCES

ABDALI, Sara; SHAHAM, Sina; KRISHNAMACHARI, Bhaskar. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. *ACM Computing Surveys*, v. 57, n. 3, p. 1–29, 2024. DOI: 10.1145/3697349.

ADEBAYO, Julius et al. Sanity Checks for Saliency Maps. In: ADVANCES in Neural Information Processing Systems 31 (NeurIPS 2018). [S.l.: s.n.], 2018. P. 9525–9536.

AÏMEUR, Esma; AMRI, Sabrine; BRASSARD, Gilles. Fake News, Disinformation and Misinformation in Social Media: A Review. *Social Network Analysis and Mining*, v. 13, n. 1, p. 30, 2023. DOI: 10.1007/s13278-023-01028-5.

ALAM, Firoj et al. A Survey on Multimodal Disinformation Detection. In: PROCEEDINGS of the 29th International Conference on Computational Linguistics, COLING 2022. [S.l.: s.n.], 2022.

ALHINDI, Tariq; PETRIDIS, Savvas; MURESAN, Smaranda. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In: PROCEEDINGS of the First Workshop on Fact Extraction and VERification (FEVER). [S.l.]: Association for Computational Linguistics, 2018. DOI: 10.18653/v1/W18-5513.

ALLAL, Loubna Ben et al. *SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model*. [S.l.: s.n.], 2025. arXiv: 2502.02737 [cs.CL]. Available from: <https://arxiv.org/abs/2502.02737>.

ALVAREZ-MELIS, David; JAAKKOLA, Tommi S. On the Robustness of Interpretability Methods. In: PROCEEDINGS of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI). [S.l.: s.n.], 2018.

AMORIM, Vicente JP et al. *Validating and monitoring microservices-based zero trust environments*. [S.l.]: Google Patents, Apr. 2025. US Patent App. 18/483,566.

ANTHROPIC. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. [S.l.: s.n.], 2024. Anthropic Model Card. Accessed: 2026-06-08. Available from: <https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf>.

ARYAL, Saugat; KEANE, Mark T. Even If Explanations: Prior Work, Desiderata & Benchmarks for Semi-Factual XAI. In: PROCEEDINGS of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. [S.l.: s.n.], 2023. P. 6526–6535. DOI: 10.24963/ijcai.2023/732.

ATANASOVA, Pepa et al. Generating Fact Checking Explanations. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020. DOI: 10.18653/v1/2020.acl-main.656.

ATREY, Pradeep K. et al. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems*, v. 16, n. 6, p. 345–379, 2010. DOI: 10.1007/s00530-010-0182-0.

AUGENSTEIN, Isabelle et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, v. 6, n. 8, p. 852–863, 2024. DOI: 10.1038/s42256-024-00881-z.

BAK-COLEMAN, Joseph B. et al. Stewardship of Global Collective Behavior. *Proceedings of the National Academy of Sciences*, v. 118, n. 27, e2025764118, 2021. DOI: 10.1073/pnas.2025764118.

BALTRUŠAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 2, p. 423–443, 2019. DOI: 10.1109/TPAMI.2018.2798607.

BEKOULIS, Giannis; PAPAGIANNOPOULOU, Christina; DELIGIANNIS, Nikos. A Review on Fact Extraction and Verification. *ACM Computing Surveys*, v. 55, n. 1, p. 1–35, 2021. DOI: 10.1145/3485127.

BIAN, Tian et al. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. In: 01. PROCEEDINGS of the AAAI Conference on Artificial Intelligence (AAAI). [S.l.: s.n.], 2020. v. 34, p. 549–556. DOI: 10.1609/aaai.v34i01.5393.

BOMMASANI, Rishi; LIANG, Percy, et al. *On the Opportunities and Risks of Foundation Models*. [S.l.: s.n.], 2021. arXiv preprint arXiv:2108.07258. Stanford Center for Research on Foundation Models (CRFM); equal contribution by Rishi Bommasani and Percy Liang. Available from: <https://arxiv.org/abs/2108.07258>.

BRODA, Elena; STRÖMBÄCK, Jesper. Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review. *Annals of the International Communication Association*, v. 48, n. 2, p. 139–166, 2024. DOI: 10.1080/23808985.2024.2323736.

BROWN, Tom B. et al. Language Models are Few-Shot Learners. In: ADVANCES in Neural Information Processing Systems 33 (NeurIPS 2020). [S.l.: s.n.], 2020. P. 1877–1901.

CANADIAN CENTRE FOR CYBER SECURITY. *How to Identify Misinformation, Disinformation, and Malinformation (ITSAP.00.300)*. [S.l.: s.n.], 2024. Government of Canada, Awareness Series. Accessed: 2026-06-06. Available from: <https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300>.

CARNEIRO, Fernando et al. BERTweet.BR: a pre-trained language model for tweets in Portuguese. *Neural Computing and Applications*, v. 37, n. 6, p. 4363–4385, 2025. DOI: 10.1007/s00521-024-10711-3.

CHANG, Yupeng et al. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, Association for Computing Machinery, v. 15, n. 3, 2024. ISSN 2157-6904. DOI: 10.1145/3641289.

CHEN, Bohan; BERTOZZI, Andrea L. AutoKG: Efficient Automated Knowledge Graph Generation for Language Models. In: IEEE. 2023 IEEE International Conference on Big Data (BigData). [S.l.: s.n.], 2023. P. 3117–3126.

CHEN, Canyu; SHU, Kai. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *AI Magazine*, v. 45, n. 3, p. 354–368, 2024. DOI: 10.1002/aaai.12188.

CHEN, Yingjian et al. GraphCheck: Breaking Long-Term Text Barriers with Extracted Knowledge Graph-Powered Fact-Checking. In: PROCEEDINGS of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2025. DOI: 10.18653/v1/2025.acl-long.729.

CHEN, Zhi-Yuan et al. Beyond the Surface: Measuring Self-Preference in LLM Judgments. In: PROCEEDINGS of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: Association for Computational Linguistics, 2025. P. 1653–1672.

CHEUNG, Tsun-Hin; LAM, Kin-Man. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. [S.l.: s.n.], 2023. DOI: 10.1109/APSIPAASC58517.2023.10317251.

CIAMPAGLIA, Giovanni Luca et al. Computational Fact Checking from Knowledge Networks. *PLOS ONE*, Public Library of Science, v. 10, n. 6, p. 1–13, June 2015. DOI: 10.1371/journal.pone.0128193. Available from: <https://doi.org/10.1371/journal.pone.0128193>.

COMITO, Carmela; CAROPRESE, Luciano; ZUMPANO, Ester. Multimodal Fake News Detection on Social Media: A Survey of Deep Learning Techniques. *Social Network Analysis and Mining*, v. 13, n. 1, p. 101, 2023. DOI: 10.1007/s13278-023-01104-w.

CRUZ FERREIRA, Victor da et al. *Recommendation framework of model and hardware for energy-efficient machine learning workloads*. [S.l.: s.n.], July 2025. US Patent 12,373,019.

CUI, Limeng; SHU, Kai, et al. DEFEND: A System for Explainable Fake News Detection. In: PROCEEDINGS of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019. [S.l.: s.n.], 2019. P. 2961–2964. DOI: 10.1145/3357384.3357862.

CUI, Wei; SHANG, Mingsheng. MIGCL: Fake News Detection with Multimodal Interaction and Graph Contrastive Learning Networks. *Applied Intelligence*, v. 55, n. 1, p. 78, 2024. DOI: 10.1007/s10489-024-05883-3.

DEVLIN, Jacob et al. 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). Minneapolis, Minnesota: Association for Computational Linguistics, 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423.

DIERICKX, Laurence; LINDÉN, Carl-Gustav; OPDAHL, Andreas Lothe. Automated Fact-Checking to Support Professional Practices: Systematic Literature Review and Meta-Analysis. *International Journal of Communication*, v. 17, p. 5170–5190, 2023.

DONG, Qingxiu et al. A Survey on In-Context Learning. In: PROCEEDINGS of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2024. DOI: 10.18653/v1/2024.emnlp-main.64.

DOSHI-VELEZ, Finale; KIM, Been. *Towards A Rigorous Science of Interpretable Machine Learning*. [S.l.: s.n.], 2017. arXiv: 1702.08608.

EILERTSEN, Brage et al. Aligning Attention with Human Rationales for Self-Explaining Hate Speech Detection. In: PROCEEDINGS of the AAAI Conference on Artificial Intelligence (AAAI). [S.l.: s.n.], 2026.

ELDIFRAWI, Islam; WANG, Shengrui; TRABELSI, Amine. Automated Justification Production for Claim Veracity in Fact Checking: A Survey on Architectures and Approaches. In: PROCEEDINGS of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.]: Association for Computational Linguistics, 2024. P. 6679–6692.

FERRAZ, Lucca Baptista Silva et al. Retrieval-Augmented Generation with Small Language Models for Fake News Detection. In: PROCEEDINGS of the 17th International Conference on Computational Processing of Portuguese (PROPOR). [S.l.]: Association for Computational Linguistics, 2026. P. 120–130.

FREUND, Werner Spolidoro et al. *Automated recommendation of most valuable solution to non-compliant cybersecurity network patterns*. [S.l.]: Google Patents, July 2025. US Patent App. 18/423,818.

FU, Jinlan et al. GPTScore: Evaluate as You Desire. In: PROCEEDINGS of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). [S.l.]: Association for Computational Linguistics, 2024. P. 6556–6576.

GAO, Mingqi et al. LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics*, v. 51, n. 2, p. 661–687, 2025. DOI: 10.1162/coli_a_00561.

GAO, Yunfan et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. [S.l.: s.n.], 2023. arXiv preprint arXiv:2312.10997. Available from: <https://arxiv.org/abs/2312.10997>.

GILMER, Justin et al. Neural Message Passing for Quantum Chemistry. In: PROCEEDINGS of the 34th International Conference on Machine Learning (ICML 2017). [S.l.: s.n.], 2017. v. 70. (Proceedings of Machine Learning Research), p. 1263–1272.

GILPIN, Leilani H. et al. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: PROCEEDINGS of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018. [S.l.: s.n.], 2018. P. 80–89. DOI: 10.1109/DSAA.2018.00018.

GONGANE, Vaishali U.; MUNOT, Mousami V.; ANUSE, Alwin D. A Survey of Explainable AI Techniques for Detection of Fake News and Hate Speech on Social Media Platforms. *Journal of Computational Social Science*, v. 7, n. 1, p. 587–623, 2024. DOI: 10.1007/s42001-024-00248-9.

GRATTAFIORI, Aaron et al. *The Llama 3 Herd of Models*. [S.l.: s.n.], 2024. arXiv preprint arXiv:2407.21783. Available from: <https://arxiv.org/abs/2407.21783>.

GU, Jiawei et al. A survey on LLM-as-a-judge. *The Innovation*, v. 7, n. 6, 2026.

GUIDOTTI, Riccardo. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, Springer, v. 38, n. 5, p. 2770–2824, 2024.

GUO, Hao; ZENG, Weixin, et al. Interpretable Fake News Detection with Graph Evidence. In: PROCEEDINGS of the 32nd ACM International Conference on Information and Knowledge Management. [S.l.: s.n.], 2023. DOI: 10.1145/3583780.3614936.

GUO, Zhijiang; SCHLICHTKRULL, M.; VLACHOS, Andreas. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, v. 10, p. 178–206, 2021.

GURRAPU, Sai; HUANG, Lifu; BATARSEH, Feras A. ExClaim: Explainable Neural Claim Verification Using Rationalization. In: 2022 IEEE 29th Annual Software Technology Conference (STC). [S.l.: s.n.], 2022. P. 19–26. DOI: 10.1109/STC55697.2022.00012.

HAIDER, Jutta; SUNDIN, Olof. *Paradoxes of media and information literacy: The crisis of information*. [S.l.]: Taylor & Francis, 2022.

HAMED, Suhaib Kh.; AB AZIZ, Mohd Juzaiddin; YAAKUB, Mohd Ridzwan. A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion. *Heliyon*, v. 9, n. 10, e20382, 2023. DOI: 10.1016/j.heliyon.2023.e20382.

HANGLOO, Sakshini; ARORA, Bhavna. Combating Multimodal Fake News on Social Media: Methods, Datasets, and Future Perspective. *Multimedia Systems*, v. 28, n. 6, p. 2391–2422, 2022. DOI: 10.1007/s00530-022-00966-y.

HATTORI, Leandro Takeshi et al. *Reliable model exchange and aggregation score in decentralized federated learning*. [S.l.]: Google Patents, May 2025. US Patent App. 18/494,625.

HOGAN, Aidan et al. Knowledge Graphs. *ACM Computing Surveys*, v. 54, n. 4, p. 1–37, 2021. DOI: 10.1145/3447772.

HOOKER, Sara et al. A Benchmark for Interpretability Methods in Deep Neural Networks. In: ADVANCES in Neural Information Processing Systems 32 (NeurIPS 2019). [S.l.: s.n.], 2019. P. 9737–9748.

HU, Beizhe et al. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In: 20. PROCEEDINGS of the AAAI Conference on Artificial Intelligence (AAAI). [S.l.: s.n.], 2024. v. 38, p. 22105–22113. DOI: 10.1609/aaai.v38i20.30214.

HU, Edward J et al. LoRA: Low-Rank Adaptation of Large Language Models. In: INTERNATIONAL Conference on Learning Representations. [S.l.: s.n.], 2022.

HU, Linmei et al. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In: PROCEEDINGS of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021. [S.l.: s.n.], 2021. P. 754–763. DOI: 10.18653/v1/2021.acl-long.62.

HU, Xuming et al. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In: PROCEEDINGS of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023. [S.l.: s.n.], 2023. P. 2901–2912. DOI: 10.1145/3539618.3591896.

HUANG, Jie; CHANG, Kevin Chen-Chuan. Towards Reasoning in Large Language Models: A Survey. In: FINDINGS of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, 2023. P. 1049–1065. DOI: 10.18653/v1/2023.findings-acl.67.

HUANG, Qiang et al. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, v. 35, n. 7, p. 6968–6972, 2023. DOI: 10.1109/TKDE.2022.3187455.

HUANG, Yani et al. *A Graph-based Verification Framework for Fact-Checking*. [S.l.: s.n.], 2025. arXiv: 2503.07282 [cs.CL]. Available from: <https://arxiv.org/abs/2503.07282>.

INTERNATIONAL FACT-CHECKING NETWORK. *International Fact-Checking Network's Code of Principles*. [S.l.: s.n.], 2023. Available from: <https://www.ifcncodeofprinciples.poynter.org/>.

JACOVI, Alon et al. Contrastive Explanations for Model Interpretability. In: PROCEEDINGS of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021. DOI: 10.18653/v1/2021.emnlp-main.120.

JI, Shaoxiong et al. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, v. 33, n. 2, p. 494–514, 2022. DOI: 10.1109/TNNLS.2021.3070843.

KAWINTIRANON, Kornraphop; SINGH, Lisa. DeMis: Data-Efficient Misinformation Detection Using Reinforcement Learning. In: MACHINE Learning and Knowledge Discovery in Databases, ECML PKDD 2023. [S.l.: s.n.], 2023. P. 224–240.

KIM, Jenia; MAATHUIS, Henry; SENT, Danielle. Human-Centered Evaluation of Explainable AI Applications: A Systematic Review. *Frontiers in Artificial Intelligence*, v. 7, 2024. DOI: 10.3389/frai.2024.1456486.

KIM, Jiho et al. FactKG: Fact Verification via Reasoning on Knowledge Graphs. In: ROGERS, Anna; BOYD-GRABER, Jordan; OKAZAKI, Naoaki (Eds.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023. P. 16190–16206. DOI: 10.18653/v1/2023.acl-long.895. Available from: <https://aclanthology.org/2023.acl-long.895/>.

KIM, Seungone et al. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. In: PROCEEDINGS of the Twelfth International Conference on Learning Representations (ICLR). [S.l.: s.n.], 2024.

KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: 3RD International Conference on Learning Representations, ICLR 2015. [S.l.: s.n.], 2015.

KIPF, Thomas N.; WELLING, Max. Semi-Supervised Classification with Graph Convolutional Networks. In: INTERNATIONAL Conference on Learning Representations (ICLR). [S.l.: s.n.], 2017.

KOKHLIKYAN, Narine et al. Captum: A Unified and Generic Model Interpretability Library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

KONG, Xiangwei; LIU, Shujie; ZHU, Luhao. Toward Human-Centered XAI in Practice: A Survey. *Machine Intelligence Research*, v. 21, n. 4, p. 740–770, 2024. DOI: 10.1007/s11633-022-1407-3.

KOTONYA, Neema; TONI, Francesca. Explainable Automated Fact-Checking for Public Health Claims. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2020. DOI: 10.18653/v1/2020.emnlp-main.623.

_____. Explainable Automated Fact-Checking: A Survey. In: PROCEEDINGS of the 28th International Conference on Computational Linguistics. [S.l.: s.n.], 2020. DOI: 10.18653/v1/2020.coling-main.474.

KOU, Ziyi; SHANG, Lanyu, et al. HC-COVID: A Hierarchical Crowdsource Knowledge Graph Approach to Explainable COVID-19 Misinformation Detection. *Proceedings of the ACM on Human-Computer Interaction*, v. 6, GROUP, 2022.

KOU, Ziyi; ZHANG, Daniel Yue, et al. ExFaux: A Weakly Supervised Approach to Explainable Fauxtography Detection. In: 2020 IEEE International Conference on Big Data, Big Data 2020. [S.l.: s.n.], 2020. P. 631–636. DOI: 10.1109/BigData50022.2020.9378019.

KRUENGKRAI, Canasai; YAMAGISHI, Junichi; WANG, Xin. A Multi-Level Attention Model for Evidence-Based Fact Checking. In: FINDINGS of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, Aug. 2021. P. 2447–2460. DOI: 10.18653/v1/2021.findings-acl.217.

LAN, Yuqing et al. Multi-Evidence Based Fact Verification via A Confidential Graph Neural Network. *IEEE Transactions on Big Data*, v. 11, n. 02, p. 426–437, 2025. ISSN 2332-7790. DOI: 10.1109/TBDATA.2024.3403382.

LAZER, David M. J. et al. The Science of Fake News. *Science*, v. 359, n. 6380, p. 1094–1096, 2018. DOI: 10.1126/science.aao2998.

LEHMANN, Jens et al. DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, v. 6, n. 2, p. 167–195, 2015. DOI: 10.3233/SW-140134.

LEWANDOWSKY, Stephan; ECKER, Ullrich K. H.; COOK, John. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, v. 6, n. 4, p. 353–369, 2017. DOI: 10.1016/j.jarmac.2017.07.008.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: ADVANCES in Neural Information Processing Systems 33 (NeurIPS 2020). [S.l.: s.n.], 2020. P. 9459–9474.

LI, Dawei et al. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In: PROCEEDINGS of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, Nov. 2025. P. 2757–2791. DOI: 10.18653/v1/2025.emnlp-main.138. Available from: <https://aclanthology.org/2025.emnlp-main.138/>.

LI, Haitao et al. *LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. [S.l.: s.n.], 2024. arXiv preprint arXiv:2412.05579. Available from: <https://arxiv.org/abs/2412.05579>.

LI, Xianghua et al. A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence*, v. 9, n. 4, p. 2658–2675, 2025. DOI: 10.1109/TETCI.2025.3543389.

LIPTON, Peter. Contrastive explanation. *Royal Institute of Philosophy Supplements*, v. 27, p. 247–266, 1990.

LIPTON, Zachary C. The Mythos of Model Interpretability. *Communications of the ACM*, v. 61, n. 10, p. 36–43, 2018. DOI: 10.1145/3233231.

LIU, Hui; WANG, Wenya; LI, Haoliang. Interpretable Multimodal Misinformation Detection with Logic Reasoning. In: FINDINGS of the Association for Computational Linguistics: ACL 2023. [S.l.]: Association for Computational Linguistics, 2023. P. 9781–9796. DOI: 10.18653/v1/2023.findings-acl.620.

LIU, Hui et al. TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. In: FINDINGS of the Association for Computational Linguistics: ACL 2024. [S.l.: s.n.], 2024. DOI: 10.18653/v1/2024.findings-acl.919.

LIU, Yang et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In: PROCEEDINGS of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: Association for Computational Linguistics, 2023. P. 2511–2522.

LOURENÇO, Vítor; FREUND, Werner, et al. *Automated online policy generation for zero-trust architectures*. [S.l.]: Google Patents, Nov. 2024. US Patent App. 18/310,804.

LOURENÇO, Vítor; PAES, Aline. A Modality-level Explainable Framework for Misinformation Checking in Social Networks. In: LATINX in AI Research at the 36th Conference on Neural Information Processing Systems, LXAI-NeurIPS 2022. [S.l.: s.n.], 2022.

_____. Learning attention-based representations from multiple patterns for relation prediction in knowledge graphs. *Knowledge-Based Systems*, Elsevier, v. 251, p. 109232, 2022.

LOURENÇO, Vítor; PAES, Aline; WEYDE, Tillman. Exploring Content and Social Connections of Fake News with Explainable Text and Graph Learning. In: PROCEEDINGS of the 35th Brazilian Conference on Intelligent Systems, BRACIS 2025. [S.l.: s.n.], 2025.

LOURENÇO, Vítor; PAES, Aline; WEYDE, Tillman, et al. KG-CRAFT: Knowledge Graph-based Contrastive Reasoning with LLMs for Enhancing Automated Fact-checking. In: PROCEEDINGS of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.]: Association for Computational Linguistics, 2026. P. 6419–6439. ISBN 979-8-89176-380-7. DOI: 10.18653/v1/2026.eacl-long.302. Available from: <https://aclanthology.org/2026.eacl-long.302/>.

LOURENÇO, Vítor N.; DUBEY, Mohnish, et al. An Explainable Natural Language Framework for Identifying and Notifying Target Audiences In Enterprise Communication. In: JOINT Proceedings of Industry, Doctoral Consortium, Posters and Demos of the 24th International Semantic Web Conference, ISWC-C 2025. [S.l.: s.n.], 2025.

LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: PROCEEDINGS of the 31st International Conference on Neural Information Processing Systems, NIPS 2017. [S.l.: s.n.], 2017. P. 4768–4777.

MACHADO, Marcelo et al. An Extensible Approach for Query-Driven Multimodal Knowledge Graph Completion. In: PROCEEDINGS of the ISWC 2022 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 21st International Semantic Web Conference, ISWC 2022. [S.l.: s.n.], 2022.

MATHEW, Binny et al. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In: 17. PROCEEDINGS of the AAAI Conference on Artificial Intelligence (AAAI). [S.l.: s.n.], 2021. v. 35, p. 14867–14875. DOI: 10.1609/aaai.v35i17.17745.

MAZURCZYK, Wojciech; LEE, Dongwon; VLACHOS, Andreas. Disinformation 2.0 in the Age of AI: A Cybersecurity Perspective. *Commun. ACM*, v. 67, n. 3, p. 36–39, 2024. ISSN 0001-0782. DOI: 10.1145/3624721.

MENG, Mark Huasong et al. *Detecting Manipulated Contents Using Knowledge-Grounded Inference*. [S.l.: s.n.], 2025. Available from: <https://arxiv.org/abs/2504.21165>.

MILLER, Tim. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, v. 267, 2019.

MIN, Sewon et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In: PROCEEDINGS of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: Association for Computational Linguistics, 2023. P. 12076–12100. DOI: 10.18653/v1/2023.emnlp-main.741.

MINH, Dang et al. Explainable Artificial Intelligence: A Comprehensive Review. *Artificial Intelligence Review*, v. 55, n. 5, p. 3503–3568, 2022. DOI: 10.1007/s10462-021-10088-y.

MOHAMMADSHAHI, Alireza et al. RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question. In: FINDINGS of the Association for Computational Linguistics: ACL 2023. [S.l.: s.n.], 2023. DOI: 10.18653/v1/2023.findings-acl.428.

MORENO, Marcio et al. A Hyperknowledge Approach to Support Dataset Engineering. In: PROCEEDINGS of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference, ISWC 2021. [S.l.: s.n.], 2021.

MORENO, Marcio et al. Evaluating Semantic Queries for Dataset Engineering on the Hyperknowledge Platform. In: PROCEEDINGS of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference, ISWC 2021. [S.l.: s.n.], 2021.

MUSEUM OF AUSTRALIAN DEMOCRACY AT OLD PARLIAMENT HOUSE. *What Is the Difference between Misinformation and Disinformation?* [S.l.: s.n.]. Museum of Australian Democracy (MoAD). Accessed: 2026-06-06. Available from: <https://moadoph.gov.au/explore/democracy/what-is-the-difference-between-misinformation-and-disinformation>.

NAKAMURA, Kai; LEVY, Sharon; WANG, William Yang. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In: PROCEEDINGS of the 12th Language Resources and Evaluation Conference, LREC 2020. [S.l.: s.n.], 2020. P. 6149–6157.

NAKOV, Preslav et al. Automated Fact-Checking for Assisting Human Fact-Checkers. In: PROCEEDINGS of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). [S.l.: s.n.], 2021. P. 4551–4558. DOI: 10.24963/ijcai.2021/619.

NAUTA, Meike et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, v. 55, 13s, p. 1–42, 2023. DOI: 10.1145/3583558.

NETO, Roberto Nery Stelling et al. *Template database generation system.* [S.l.]: Google Patents, July 2025. US Patent App. 18/423,140.

NEWMAN, Nic et al. *Reuters Institute Digital News Report 2025.* [S.l.], 2025.

NGUEAJIO, Mikel K. et al. Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *ACM Computing Surveys*, v. 57, n. 7, p. 1–37, 2025. DOI: 10.1145/3711123.

NGUYEN, Dat Quoc; VU, Thanh; TUAN NGUYEN, Anh. BERTweet: A Pre-trained Language Model for English Tweets. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. [S.l.: s.n.], 2020. P. 9–14. DOI: 10.18653/v1/2020.emnlp-demos.2.

NIELSEN, Dan S.; MCCONVILLE, Ryan. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In: PROCEEDINGS of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022. [S.l.: s.n.], 2022. P. 3141–3153. DOI: 10.1145/3477495.3531744.

PAN, Liangming et al. Fact-Checking Complex Claims with Program-Guided Reasoning. In: PROCEEDINGS of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.]: Association for Computational Linguistics, 2023. P. 6981–7004. DOI: 10.18653/v1/2023.acl-long.386.

PAN, Shirui et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, v. 36, n. 7, p. 3580–3599, 2024. DOI: 10.1109/TKDE.2024.3352100.

PANICKSSERY, Arjun; BOWMAN, Samuel R.; FENG, Shi. LLM Evaluators Recognize and Favor Their Own Generations. In: ADVANCES in Neural Information Processing Systems 37 (NeurIPS 2024). [S.l.: s.n.], 2024.

PARANJAPE, Bhargavi et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. In: FINDINGS of the Association for Computational Linguistics: ACL-IJCNLP 2021. [S.l.: s.n.], 2021. DOI: 10.18653/v1/2021.findings-acl.366.

PASZKE, Adam et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: ADVANCES in Neural Information Processing Systems, NeurIPS 2019. [S.l.: s.n.], 2019. P. 8024–8035.

PENG, Boci et al. Graph Retrieval-Augmented Generation: A Survey. *ACM Transactions on Information Systems*, v. 44, n. 2, p. 1–52, 2025. DOI: 10.1145/3777378.

PENNYCOOK, Gordon; RAND, David G. The Psychology of Fake News. *Trends in Cognitive Sciences*, v. 25, n. 5, p. 388–402, 2021. DOI: 10.1016/j.tics.2021.02.007.

PÉREZ, Juan Manuel et al. RoBERTuito: A Pre-trained Language Model for Social Media Text in Spanish. In: PROCEEDINGS of the Language Resources and Evaluation Conference, LREC 2022. [S.l.: s.n.], 2022. P. 7235–7243.

PEW RESEARCH CENTER. *Social Media and News Fact Sheet*. [S.l.: s.n.], 2025. Pew Research Center. Accessed: 2026-06-21. Available from: <https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/>.

POMBAL, José; REI, Ricardo; MARTINS, André F. T. *Self-Preference Bias in Rubric-Based Evaluation of Large Language Models*. [S.l.: s.n.], 2026. arXiv preprint arXiv:2604.06996. Preprint, under review. Available from: <https://arxiv.org/abs/2604.06996>.

QI, P. et al. Exploiting Multi-domain Visual Information for Fake News Detection. In: 2019 IEEE International Conference on Data Mining, ICDM 2019. [S.l.: s.n.], 2019. P. 518–527. DOI: 10.1109/ICDM.2019.00062.

QIAN, Shengsheng; HU, Jun, et al. Knowledge-Aware Multi-Modal Adaptive Graph Convolutional Networks for Fake News Detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, v. 17, n. 3, p. 1–23, 2021. DOI: 10.1145/3451215.

QIAN, Shengsheng; WANG, Jinguang, et al. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In: PROCEEDINGS of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021. [S.l.: s.n.], 2021. P. 153–162. DOI: 10.1145/3404835.3462871.

QUDUS, Umair et al. Fact Checking Knowledge Graphs – A Survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, July 2025. ISSN 0360-0300. DOI: 10.1145/3749838. Available from: <https://doi.org/10.1145/3749838>.

RANGEL, Rodrigo Fill et al. A Unified Framework for Cropland Field Boundary Detection and Segmentation. In: PROCEEDINGS of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024. [S.l.: s.n.], 2024. P. 636–644. DOI: 10.1109/WACVW60836.2024.00073.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016. [S.l.: s.n.], 2016. P. 1135–1144. DOI: 10.1145/2939672.2939778.

RONG, Yao et al. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, n. 4, p. 2104–2122, 2024. DOI: 10.1109/TPAMI.2023.3331846.

RUSSO, Daniel; TEKIROĞLU, Serra Sinem; GUERINI, Marco. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics*, v. 11, p. 1250–1264, 2023. DOI: 10.1162/tacl_a_00601.

SAEED, Waddah; OMLIN, Christian. Explainable AI (XAI): A Systematic Meta-survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems*, v. 263, p. 110273, 2023. DOI: 10.1016/j.knosys.2023.110273.

SAHNAN, Dhruv et al. Can LLMs Automate Fact-Checking Article Writing? *Transactions of the Association for Computational Linguistics*, v. 14, p. 489–509, 2026. DOI: 10.1162/tacl.a.644.

SALLES, Isadora; VARGAS, Francielle; BENEVENUTO, Fabrício. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. In: PROCEEDINGS of the 31st International Conference on Computational Linguistics (COLING 2025). Abu Dhabi, UAE: Association for Computational Linguistics, 2025. P. 6659–6669.

SCARSELLI, Franco et al. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, v. 20, n. 1, p. 61–80, 2009. DOI: 10.1109/TNN.2008.2005605.

SCHNEIDER, Johannes. Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda. *Artificial Intelligence Review*, v. 57, n. 11, 2024. DOI: 10.1007/s10462-024-10916-x.

SHANG, Lanyu et al. A Duo-generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. *Proceedings of the ACM Web Conference, WWW 2022*, p. 3623–3631, 2022.

SHIRALKAR, Prashant et al. Finding Streams in Knowledge Graphs to Support Fact Checking. In: 2017 IEEE International Conference on Data Mining (ICDM). [S.l.: s.n.], 2017. P. 859–864. DOI: 10.1109/ICDM.2017.105.

SHU, Kai; CUI, Limeng, et al. dEFEND: Explainable Fake News Detection. In: PROCEEDINGS of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [S.l.: s.n.], 2019. DOI: 10.1145/3292500.3330935.

SHU, Kai; MAHUDESWARAN, Deepak, et al. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. In: INTERNATIONAL Conference on Web and Social Media, ICWSM 2020. [S.l.: s.n.], 2020.

SHU, Kai; SLIVA, Amy, et al. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, v. 19, n. 1, p. 22–36, 2017.

SHU, Kai; WANG, Suhang; LIU, Huan. Beyond News Contents: The Role of Social Context for Fake News Detection. In: PROCEEDINGS of the 12th ACM International Conference on Web Search and Data Mining, WSDM 2019. [S.l.: s.n.], 2019. P. 312–320. DOI: 10.1145/3289600.3290994.

SINGAL, Ronit et al. Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. In: PROCEEDINGS of the Seventh Fact Extraction and VERification Workshop. [S.l.: s.n.], 2024. DOI: 10.18653/v1/2024.fever-1.10.

SINGH, Bhuvanesh; SHARMA, Dilip Kumar. SiteForge: Detecting and Localizing Forged Images on Microblogging Platforms Using Deep Convolutional Neural Network. *Computers & Industrial Engineering*, v. 162, 2021. DOI: 10.1016/j.cie.2021.107733.

SINGHAL, Shivangi; SHAH, Rajiv Ratn; KUMARAGURU, Ponnurangam. FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India. *Proceedings of the International AAAI Conference on Web and Social Media, ICWSM 2022*, v. 16, n. 1, p. 1322–1331, 2022.

SOUZA, Renan et al. Workflow Provenance in the Lifecycle of Scientific Machine Learning. *Concurrency Computation Practice and Experience*, Wiley, p. 1–21, 2021.

STEPIN, Ilia et al. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, v. 9, 2021. DOI: 10.1109/ACCESS.2021.3051315.

SUNDARARAJAN, Mukund; TALY, Ankur; YAN, Qiqi. Axiomatic Attribution for Deep Networks. In: PROCEEDINGS of the 34th International Conference on Machine Learning, ICML 2017. [S.l.: s.n.], 2017. P. 3319–3328.

TAN, Xin; ZOU, Bowei; AW, Ai Ti. Improving Explainable Fact-Checking with Claim-Evidence Correlations. In: PROCEEDINGS of the 31st International Conference on Computational Linguistics. [S.l.: s.n.], 2025.

TANG, Liyan; LABAN, Philippe; DURRETT, Greg. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In: PROCEEDINGS of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: Association for Computational Linguistics, 2024. P. 8818–8847. DOI: 10.18653/v1/2024.emnlp-main.499.

TEWKSBURY, David; RITTENBERG, Jason. *News on the Internet: Information and Citizenship in the 21st Century*. [S.l.]: Oxford University Press, 2012. DOI: 10.1093/acprof:osobl/9780195391961.001.0001.

THORNE, James; VLACHOS, Andreas. Automated Fact Checking: Task Formulations, Methods and Future Directions. In: PROCEEDINGS of the 27th International Conference on Computational Linguistics, COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018. P. 3346–3359.

THORNE, James; VLACHOS, Andreas, et al. 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). [S.l.: s.n.], 2018. DOI: 10.18653/v1/N18-1074.

TRAGER, Jackson et al. MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of LLMs through Multi-hop Hate Speech Explanation. In: FINDINGS of the Association for Computational Linguistics: EMNLP 2025. [S.l.]: Association for Computational Linguistics, 2025. P. 15709–15740.

TUFCHI, Shivani; YADAV, Ashima; AHMED, Tanveer. A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities. *International Journal of Multimedia Information Retrieval*, v. 12, n. 2, p. 28, 2023. DOI: 10.1007/s13735-023-00296-3.

VALENZUELA, Sebastián et al. The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, Taylor & Francis, v. 7, n. 6, p. 802–823, 2019.

VARGAS, Francielle; BENEVENUTO, Fabrício; PARDO, Thiago A. S. Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection. In: PROCEEDINGS of the 17th International Conference on Computational Processing of Portuguese (PROPOR). [S.l.]: Association for Computational Linguistics, 2026. P. 35–42.

VARGAS, Francielle; SALLES, Isadora, et al. Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning. In: PROCEEDINGS of the Seventh Fact Extraction and VERification Workshop (FEVER). [S.l.]: Association for Computational Linguistics, 2024. P. 192–204.

VASWANI, Ashish et al. Attention is all you need. In: PROCEEDINGS of the 31st International Conference on Neural Information Processing Systems. [S.l.: s.n.], 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VELIČKOVIĆ, Petar et al. Graph Attention Networks. In: INTERNATIONAL Conference on Learning Representations, ICLR 2018. [S.l.: s.n.], 2018.

VERGA, Pat et al. *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*. [S.l.: s.n.], 2024. arXiv preprint arXiv:2404.18796. Available from: <https://arxiv.org/abs/2404.18796>.

VERMA, Pawan Kumar et al. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*, v. 8, n. 4, p. 881–893, 2021. DOI: 10.1109/TCSS.2021.3068519.

VERMA, Sahil et al. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, ACM New York, NY, v. 56, n. 12, p. 1–42, 2024.

VLACHOS, Andreas; RIEDEL, Sebastian. Fact Checking: Task Definition and Dataset Construction. In: PROCEEDINGS of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, MD, USA: Association for Computational Linguistics, June 2014. P. 18–22. DOI: 10.3115/v1/W14-2508.

VLADIKA, Juraj; SCHNEIDER, Phillip; MATTHES, Florian. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. In: PROCEEDINGS of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy: [s.n.], 2024. P. 8095–8107.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The Spread of True and False News Online. *Science*, v. 359, n. 6380, p. 1146–1151, 2018. DOI: 10.1126/science.aap9559.

VRANDEČIĆ, Denny; KRÖTZSCH, Markus. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, v. 57, n. 10, p. 78–85, 2014. DOI: 10.1145/2629489.

WACHTER, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, v. 31, n. 2, p. 841–887, 2017. DOI: 10.2139/ssrn.3063289.

WADDEN, David; LIN, Shanchuan, et al. Fact or Fiction: Verifying Scientific Claims. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020. P. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609.

WADDEN, David; LO, Kyle, et al. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In: FINDINGS of the Association for Computational Linguistics: NAACL 2022. [S.l.: s.n.], 2022. DOI: 10.18653/v1/2022.findings-naacl.6.

WANG, Bo et al. Explainable Fake News Detection with Large Language Model via Defense Among Competing Wisdom. In: PROCEEDINGS of the ACM Web Conference 2024. [S.l.: s.n.], 2024. DOI: 10.1145/3589334.3645471.

WANG, Jinguang et al. End-to-End Explainable Fake News Detection Via Evidence-Claim Variational Causal Inference. *ACM Trans. Inf. Syst.*, v. 43, n. 4, 2025. ISSN 1046-8188. DOI: 10.1145/3728462.

WANG, Peiyi et al. Large Language Models are not Fair Evaluators. In: PROCEEDINGS of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.]: Association for Computational Linguistics, 2024. P. 9440–9450.

WANG, William Yang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: BARZILAY, Regina; KAN, Min-Yen (Eds.). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017. P. 422–426. DOI: 10.18653/v1/P17-2067. Available from: <https://aclanthology.org/P17-2067/>.

WANG, Y. et al. Understanding the Use of Fauxtography on Social Media. In: THE 15th International AAAI Conference on Web and Social Media, ICWSM 2021. [S.l.: s.n.], 2021.

WANG, Yuxia et al. Factuality of Large Language Models: A Survey. In_____. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024. P. 19519–19529. DOI: 10.18653/v1/2024.emnlp-main.1088. Available from: <https://aclanthology.org/2024.emnlp-main.1088/>.

WARDLE, Claire. *Fake News. It's Complicated.* [S.l.: s.n.], Feb. 2017. First Draft News. Accessed: 2026-06-06. Available from: <https://firstdraftnews.org/articles/fake-news-complicated/>.

WARDLE, Claire; DERAKHSHAN, Hossein. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Strasbourg, 2017.

WARREN, Greta; SHKLOVSKI, Irina; AUGENSTEIN, Isabelle. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. In: PROCEEDINGS of the 2025 CHI Conference on Human Factors in Computing Systems. [S.l.: s.n.], 2025. DOI: 10.1145/3706598.3713277.

WEI, Jason et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: ADVANCES in Neural Information Processing Systems 35 (NeurIPS 2022). [S.l.: s.n.], 2022. P. 24824–24837.

WEI, Jerry et al. Long-form Factuality in Large Language Models. In: ADVANCES in Neural Information Processing Systems 37 (NeurIPS 2024). [S.l.: s.n.], 2024.

WHITEHOUSE, Chenxi et al. Evaluation of Fake News Detection with Knowledge-Enhanced Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 16, n. 1, 2022. DOI: 10.1609/icwsm.v16i1.19400.

WU, Zonghan et al. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, v. 32, n. 1, p. 4–24, 2021. DOI: 10.1109/TNNLS.2020.2978386.

XIE, Zhuohan et al. FIRE: Fact-checking with Iterative Retrieval and Verification. In: FINDINGS of the Association for Computational Linguistics: NAACL 2025. [S.l.: s.n.], 2025. DOI: 10.18653/v1/2025.findings-naacl.158.

XIONG, Cheng et al. DelphiAgent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing & Management*, v. 62, n. 6, p. 104241, 2025. ISSN 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2025.104241.

YAMADA, Makoto et al. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, v. 26, n. 1, p. 185–207, 2014. DOI: 10.1162/NECO_a_00537.

YANG, An et al. *Qwen3 Technical Report*. [S.l.: s.n.], 2025. arXiv: 2505.09388 [cs.CL]. Available from: <https://arxiv.org/abs/2505.09388>.

YANG, Shuo et al. RealFactBench: A Benchmark for Evaluating Large Language Models in Real-World Fact-Checking. In: PROCEEDINGS of the 33rd ACM International Conference on Multimedia (MM). [S.l.]: Association for Computing Machinery, 2025. DOI: 10.1145/3746027.3758307.

YANG, Zhiwei et al. A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. In: PROCEEDINGS of the 29th International Conference on Computational Linguistics. [S.l.: s.n.], 2022.

YAO, Barry Menglong et al. 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 2023. [S.l.: s.n.], 2023. P. 2733–2743. DOI: 10.1145/3539618.3591879.

YE, Seonghyeon et al. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In: PROCEEDINGS of the Twelfth International Conference on Learning Representations (ICLR). [S.l.: s.n.], 2024.

YUE, Zhenrui et al. Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments. In: PROCEEDINGS of the 62nd Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2024. DOI: 10.18653/v1/2024.acl-long.556.

ZENG, Fengzhu; GAO, Wei. JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, v. 12, p. 334–354, 2024. DOI: 10.1162/tacl_a_00649.

ZHA, Yuheng et al. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In: PROCEEDINGS of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2023. Available from: <https://aclanthology.org/2023.acl-long.634>.

ZHANG, Bowen; SOH, Harold. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In: PROCEEDINGS of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2024. P. 9820–9836. DOI: 10.18653/v1/2024.emnlp-main.548.

ZHANG, Xuan; GAO, Wei. Reinforcement Retrieval Leveraging Fine-grained Feedback for Fact Checking News Claims with Black-Box LLM. In: PROCEEDINGS of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. [S.l.: s.n.], 2024.

ZHAO, Xiaoyan et al. PACAR: Automated Fact-Checking with Planning and Customized Action Reasoning Using Large Language Models. In: PROCEEDINGS of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). [S.l.: s.n.], 2024. P. 12564–12573.

ZHENG, Lianmin et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: ADVANCES in Neural Information Processing Systems 36 (NeurIPS 2023). [S.l.: s.n.], 2023.

ZHOU, Xinyi; MULAY, Apurva, et al. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In: PROCEEDINGS of the 29th ACM International Conference on Information & Knowledge Management, CIKM 2020. [S.l.: s.n.], 2020. P. 3205–3212. DOI: 10.1145/3340531.3412880.

ZHOU, Xinyi; WU, Jindi; ZAFARANI, Reza. SAFE: Similarity-Aware Multi-modal Fake News Detection. In: ADVANCES in Knowledge Discovery and Data Mining, PAKDD 2020. [S.l.: s.n.], 2020. P. 354–367.

ZHOU, Xinyi; ZAFARANI, Reza. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys (CSUR)*, v. 53, n. 5, 2020. DOI: 10.1145/3395046.

ZHU, Yuqi et al. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, Springer, v. 27, n. 5, p. 58, 2024.

APPENDIX A – Prompt Templates

This appendix presents the verbatim text of every LLM prompt used in the contributions of Chapter 4 (KG-CRAFT pipeline) and Chapter 6 (justification generation and evaluation). The prompts are reproduced as deployed in our experiments, with placeholders enclosed in braces (*e.g.*, {claim}, {context}) marking instance-specific fields filled in at runtime.

A.1 Chapter 4: KG-CRAFT Pipeline Prompts

This section presents the five prompts of Chapter 4 used in the scope of this work: Knowledge Graph Extraction (section 4.3.2), Contrastive Question Answer Generation (section 4.3.3), Answer Summarisation (section 4.3.3), Claim Veracity Verification (section 4.3.4), and Contrastive Question Formulation (section 4.5.2.1). For each prompt, we provide the template.

A.1.1 Knowledge Graph Extraction Prompt

The following prompt describes our phased approach to knowledge graph extraction, where the LLM is guided to sequentially identify entities ($\mathcal{E}$), assign their classes ($\mathbb{C}$), and establish relationships ($\mathcal{R}$) between them to construct the input knowledge graph $\mathcal{G}$.

Knowledge Graph Extraction Prompt

```
You are a top-tier algorithm designed for extracting information in structured
formats to build a knowledge graph.
Knowledge graphs consist of a set of triples. Each triple contains two entities
(subject and object) and one relation that connects these subject and object.
Try to capture as much information from the text as possible without sacrificing
accuracy. Do not add any information that is not explicitly mentioned in the text.
This is the process to extract information and build a knowledge graph:
1. Extract nodes [...]
2. Label nodes [...]
```

```
3. Extract relationships [...]
Compliance criteria: [...]
Text: {claim or report}
```

A.1.2 Contrastive Question Answer Generation Prompt

The following prompt instructs the LLM to generate answers to contrastive questions ($\mathcal{Q}^{\mathsf{K}}_{ranked}$) by analysing claim-associated reports ($\mathcal{R}_{\mathcal{C}}$), ensuring responses are grounded in evidence whilst maintaining traceability between claims, reports, and questions.

Contrastive Question Answer Generation Prompt

```
You are an expert answering questions based only on the provided context.

## Task:
Using the context provided and being aware of the claim, answer the question
regarding the claim aiming to fact-check it. Limit your answer to 200 words at most.

## Desired Outcome:
- Base the concise answer strictly on the context.
- Present the information neutrally, without judging or labeling the claim.
- Do not re-state the claim in the answer.
- Write in continuous prose (no lists, bullet points, or meta-commentary).
- Limit your answer to 200 words at most.

## Input:
* Context: {context}
* Claim: {claim}
* Question: {contrastive question}

## Output:
```

A.1.3 Answer Summarisation Prompt

The following prompt instructs the LLM to generate a concise summary ($A_{\mathcal{C}}$) from the claim ($\mathcal{C}$) and its associated question-answer pairs ($\mathcal{Q}^{\mathsf{K}}_{ranked}$, $\tilde{\mathcal{A}}$), emphasizing key contrasting elements while abstracting non-essential information.

Answer Summarisation Prompt

```
You are an expert writing summarizing information from pairs of question and answer.

## Task:
Your task is to generate an one paragraph summary of the information based on given
pairs of question and answer.

## Desired Outcome:
- A one paragraph summary of the information contained in the question and answer.
- Present the information neutrally, without judging or labeling the claim.
- Ensure that the summary is clear and accurately based on the provided context.
- Write in continuous prose (no lists, bullet points, or meta-commentary).
- Do not add any utterances (for example "Here are" statements) to the final answer.
- Limit your answer to 200 words at most.

## Input:
* Question 1: {contrastive question}
* Answer 1: {contrastive question answer}
* [...]

## Output:
```

A.1.4 Claim Veracity Verification Prompt

The following prompt (p_{cv}) instructs the LLM to determine claim veracity ($\mathcal{V}_{\mathcal{C}}$) by analysing the original claim ($\mathcal{C}$) against the distilled evidence summary ($A_{\mathcal{C}}$), using predefined veracity labels and their descriptions.

Claim Veracity Verification Prompt

```
You are an expert fact-checking claims based solely on the provided context of the
claim.

## Task:
Your task is to categorize the claim based only on the context as:
- {veracity labels}

## Desired Outcome:
- The veracity of the claim based on the context provided.
- Your response must be only one of the he above options. Do not include any other
text.
```

```
## Input:
* Context: {context}
* Claim: {claim}

## Output:
```

A.1.5 Contrastive Question Formulation Prompt

The following prompt guides the LLM to generate contrastive questions directly from claims and reports, serving as an alternative to the knowledge graph-based approach for our ablation study.

Contrastive Question Formulation Prompt

```
You are an expert writing analyzing a given claim and generating contrastive
questions based on given context. Your task is to generate contrastive questions of
given claim based on given context.

## Desired Outcome:
- Create contrastive questions of an input claim based on given context.
- Present a list of five contrastive questions.
- Do not add any utterances (for example "Here are" statements) to the final answer.

## Example Prompt:
* Claim: {claim example}
* Context: {reports examples}

## Example Output:
"""
{contrastive questions examples}
"""

## Additional Notes:
- Ensure that the contrastive questions are clear and accurately contrasts the
claim based on the provided context.
- Maintain a consistent and readable format for the output.
- Ensure that the output is only the contrastive questions, no other additional
text or utterances.

## Input:
* Claim: {claim}
```

```
* Context: {reports}

## Output:
```

A.2 Chapter 6: Justification Generation and Evaluation Prompts

This section presents the seven prompts used in Chapter 6: the justification generation prompt p_{just} (Section 6.3.2), and the six desideratum evaluation prompts p_d (Section 6.3.3) corresponding to *Actionable*, *Causal*, *Contextful*, *Impartial*, *Parsimonious*, and *Safety*. For each prompt, we provide the verbatim template; placeholders enclosed in braces (*e.g.*, {claim}, {label}, {context}) are filled in at runtime per fact-checking instance.

A.2.1 Justification Generation Prompt

The justification prompt p_{just} takes the claim, the predicted veracity label, and a system-specific context (Section 6.3.2) and returns a natural language justification. The prompt is identical across all four LLM-based methods compared in this chapter; only the context argument varies.

Justification Generation — System

```
You are an expert fact-checker specialized in analyzing claims and providing
justifications.

Your task is to generate justifications explaining why claims have been classified
with specific veracity labels.

### Output Requirements:
- Provide a coherent justification paragraph
- Limit responses to 200 words maximum
```

Justification Generation — User

```
**Claim:** {claim}

**Predicted Veracity:** {label}

**Supporting Context:**
{context}

Based on the context provided, explain why the claim has been classified as {label}.

**Justification:**
```

A.2.2 Actionability Evaluation Prompt

The actionable evaluation prompt $p_{actionable}$ instructs the LLM-as-a-Judge to score a candidate justification along the Actionable desideratum (Definition 6.1) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

Actionability Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **actionability** of
fact-checking justifications. You will evaluate whether a given justification
provides actionable guidance that users can follow to understand and address
misinformation.

## Definition: Actionable Explanation
An actionable explanation provides users with concrete guidelines indicating steps
towards a desired outcome. In fact-checking contexts, actionable explanations
should:
- Provide factual corrections for erroneous claims
- Indicate which specific parts of a claim contain misinformation
- Offer corrected versions of false statements
- Identify credible sources that counter the misinformation
- Flag repeated falsehoods when applicable
- Provide contextual data to support corrections

## Evaluation Scale (1-5 Likert Scale)

**Score 1 - Not Actionable**
- No factual corrections provided
- Fails to identify specific misinformation elements
- No credible sources cited
- Provides only generic statements without actionable guidance
- Does not help users understand how to address the misinformation

**Score 2 - Minimally Actionable**
- Very limited factual corrections
- Vague identification of problematic elements
- Minimal or unreliable source citations
- Provides some guidance but lacks specificity and clarity
- Limited utility for addressing misinformation

**Score 3 - Moderately Actionable**
- Some factual corrections provided
- Partially identifies specific misinformation elements
- Includes some credible sources but may be incomplete
- Offers moderate guidance with room for improvement
```

```
- Provides reasonable but not comprehensive actionable steps

**Score 4 - Mostly Actionable**
- Clear factual corrections for most erroneous elements
- Effectively identifies specific parts containing misinformation
- Cites multiple credible sources
- Provides clear, specific guidance for most aspects
- Minor gaps in actionability but overall strong performance

**Score 5 - Fully Actionable**
- Comprehensive factual corrections for all erroneous claims
- Precisely identifies all misinformation elements
- Provides multiple high-quality, credible sources
- Offers clear, specific, and complete guidance
- Enables users to fully understand and address the misinformation

## Instructions
Evaluate the given fact-checking justification for its actionability.

Based on the inputs, assess how well the justification demonstrates actionable
guidance by:
1. **Analyzing** the presence and quality of factual corrections
2. **Assessing** how clearly it identifies specific misinformation elements
3. **Evaluating** the credibility and relevance of cited sources
4. **Determining** the overall utility for addressing misinformation

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation of your scoring decision, highlighting strengths
and weaknesses in actionability]",
"dimension": "actionable"
}

## Input
```

A.2.3 Causality Evaluation Prompt

The causal evaluation prompt p_{causal} instructs the LLM-as-a-Judge to score a candidate justification along the Causal desideratum (Definition 6.2) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

Causality Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **causal reasoning** quality
of fact-checking justifications. You will evaluate whether a given justification
demonstrates causal relationships between inputs and predictions, and provides
causal explanations for the fact-checking decision.

## Definition: Causal Explanation
A causal explanation demonstrates the use of causal models to infer causal
relationships between inputs and outputted predictions. In fact-checking contexts,
causal explanations should:
- Establish clear causal relationships between evidence and conclusions
- Provide counterfactual explanations (e.g., "If X were different, then Y would
change")
- Demonstrate how specific evidence causally leads to the fact-checking verdict
- Show causal chains of reasoning rather than mere correlational observations
- Explain why certain factors are causally relevant to the truth/falsity
determination
- Use causal language and reasoning patterns

## Evaluation Scale (1-5 Likert Scale)

**Score 1 - No Causal Reasoning**
- No evidence of causal thinking or causal relationships
- Purely descriptive without causal connections
- No counterfactual reasoning provided
- Fails to explain why evidence leads to conclusions
- Uses only correlational or associative language

**Score 2 - Minimal Causal Reasoning**
- Very limited causal connections suggested
- Weak or unclear causal relationships
- No meaningful counterfactual explanations
- Minimal explanation of causal pathways
- Mostly descriptive with occasional causal hints

**Score 3 - Moderate Causal Reasoning**
- Some causal relationships established
- Basic counterfactual reasoning present
- Partial explanation of how evidence causes conclusions
- Inconsistent use of causal reasoning
- Mix of causal and non-causal explanations

**Score 4 - Strong Causal Reasoning**
- Clear causal relationships between evidence and conclusions
- Good use of counterfactual explanations
```

```
- Mostly demonstrates causal pathways
- Consistent causal reasoning with minor gaps
- Strong causal language and logical structure

**Score 5 - Excellent Causal Reasoning**
- Comprehensive causal model evident
- Rich counterfactual explanations throughout
- Clear causal chains from evidence to verdict
- Sophisticated causal reasoning and inference
- Demonstrates why alternatives would lead to different conclusions

## Instructions
Evaluate the given fact-checking justification for its causal reasoning quality.

Based on the inputs, assess how well the justification demonstrates causal
reasoning by:
1. **Analyzing** the presence of causal relationships between evidence and
conclusions
2. **Assessing** the quality and clarity of counterfactual explanations
3. **Evaluating** the use of causal language and reasoning patterns
4. **Determining** how well the justification explains causal pathways to the
verdict

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation of your scoring decision, highlighting strengths
and weaknesses in actionability]",
"dimension": "causal"
}

## Input
```

A.2.4 Contextfulness Evaluation Prompt

The contextful evaluation prompt $p_{contextful}$ instructs the LLM-as-a-Judge to score a candidate justification along the Contextful desideratum (Definition 6.3) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

Contextfulness Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **context-fullness** of
fact-checking justifications. You will evaluate whether a given justification is
presented in a way that allows its context to be fully understood by readers.
```

```
## Definition: Context-Full Explanation
A context-full explanation is one that is presented in a way that its context can
be fully understood. In fact-checking contexts, context-full explanations should:
- Be self-contained and comprehensible within the provided inputs
- Effectively utilize and reference the broader context (claim, background
information, etc.)
- Allow readers to fully understand the situational context without requiring
external information
- Clearly situate the explanation within the given claim and contextual framework
- Provide sufficient background and contextual information for complete
understanding
- Connect the justification meaningfully to the original claim and provided context

## Evaluation Scale (1-5 Likert Scale)

**Score 1 - Not Context-Full**
- Explanation lacks contextual grounding
- Difficult to understand without external information
- Poor integration with provided claim and context
- Missing essential contextual elements
- Explanation appears disconnected from inputs

**Score 2 - Minimally Context-Full**
- Limited contextual integration
- Some references to claim/context but insufficient for full understanding
- Requires additional external knowledge to comprehend
- Weak connection between explanation and provided inputs
- Incomplete contextual framework

**Score 3 - Moderately Context-Full**
- Adequate use of provided context
- Generally understandable within given framework
- Some gaps in contextual integration
- Mostly self-contained but may benefit from additional context
- Reasonable connection to claim and background information

**Score 4 - Mostly Context-Full**
- Good integration of claim and contextual information
- Largely self-contained and understandable
- Effective use of provided inputs for contextualization
- Minor gaps in contextual completeness
- Strong connection between explanation and given context

**Score 5 - Fully Context-Full**
- Excellent integration of all provided contextual information
- Completely self-contained and fully understandable
```

```
- Masterful use of claim and background context
- No external information needed for comprehension
- Perfect situational grounding within provided inputs

## Instructions
Evaluate the given fact-checking justification for its context-fullness.

Based on the inputs, assess how well the justification utilizes and integrates the
provided context by:
1. **Analyzing** how well the explanation is situated within the given claim and
context
2. **Assessing** whether the justification is self-contained and understandable
3. **Evaluating** the integration of contextual information throughout the
explanation
4. **Determining** whether additional external information is needed for full
understanding

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation of your scoring decision, highlighting how well
the justification integrates contextual information and whether it can be fully
understood within the provided framework]",
"dimension": "context_full"
}

## Input
```

A.2.5 Impartiality Evaluation Prompt

The impartial evaluation prompt $p_{impartial}$ instructs the LLM-as-a-Judge to score a candidate justification along the Impartial desideratum (Definition 6.4) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

Impartiality Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **impartiality** of
fact-checking justifications. You will evaluate whether a given justification is
free from bias and avoids the influence of partisan language, opinions, or biased
evidence sources.

## Definition: Impartial Explanation
An unbiased or impartial explanation is one for which biases in the evidence data
and other inputs have not influenced the explanation. In fact-checking contexts,
```

```
impartial explanations should:
- Avoid hyper-partisan language or politically charged rhetoric
- Present factual information without ideological bias
- Distinguish clearly between facts and opinions
- Use neutral, objective language throughout
- Rely on credible, non-partisan sources
- Avoid loaded terms, inflammatory language, or emotional appeals
- Present information in a balanced manner without favoring particular political
perspectives
- Focus on verifiable facts rather than subjective interpretations

## Evaluation Scale (1-5 Likert Scale)

**Score 1 - Highly Biased/Partial**
- Strong presence of partisan language or ideological bias
- Opinions presented as facts without distinction
- Heavy use of loaded, inflammatory, or emotionally charged terms
- Clear political or ideological slant throughout
- Relies heavily on biased or partisan sources

**Score 2 - Moderately Biased/Partial**
- Noticeable partisan language or bias in several areas
- Some mixing of opinions with factual claims
- Occasional use of loaded terms or charged language
- Identifiable ideological leaning in presentation
- Mix of biased and neutral sources

**Score 3 - Somewhat Neutral/Impartial**
- Generally neutral with occasional bias indicators
- Mostly factual with minor opinion intrusions
- Limited use of charged language
- Attempts neutrality but with some lapses
- Predominantly reliable sources with some potential bias

**Score 4 - Mostly Impartial**
- Largely neutral and objective presentation
- Clear distinction between facts and opinions
- Minimal use of charged or partisan language
- Professional, balanced tone throughout
- Reliance on credible, mostly neutral sources

**Score 5 - Fully Impartial**
- Completely neutral and objective presentation
- Strict separation of facts from opinions
- No partisan language or ideological bias detected
- Professional, balanced, and factual tone
```

```
- Exclusive use of credible, neutral, and reliable sources

## Instructions
Evaluate the given fact-checking justification for its impartiality.

Based on the inputs, assess how well the justification maintains impartiality by:
1. **Analyzing** the presence of partisan language, loaded terms, or ideological
bias
2. **Assessing** whether opinions are clearly distinguished from factual statements
3. **Evaluating** the neutrality and objectivity of the language used
4. **Determining** whether the explanation relies on credible, non-partisan sources

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation of your scoring decision, highlighting any
detected bias, partisan language, or neutral presentation, and noting the
objectivity of sources and language used]",
"dimension": "impartial"
}

## Input
```

A.2.6 Parsimony Evaluation Prompt

The parsimonious evaluation prompt $p_{parsimonious}$ instructs the LLM-as-a-Judge to score a candidate justification along the Parsimonious desideratum (Definition 6.5) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

Parsimony Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **parsimony** of fact-checking
justifications. You will evaluate whether a given justification communicates all
necessary information with minimal redundant text and maximum brevity.

## Definition: Parsimonious Explanation
A parsimonious explanation relates to brevity and efficiency in communication. It
should communicate all information needed with minimal redundant text. In
fact-checking contexts, parsimonious explanations should:
- Be concise and to-the-point without unnecessary elaboration
- Avoid redundant statements, repetitive information, or circular reasoning
- Communicate essential information efficiently and clearly
- Eliminate verbose or wordy passages that don't add value
- Use direct, clear language without excessive qualification or hedging
```

```
- Focus on key evidence and reasoning without tangential details
- Maintain completeness while maximizing brevity
- Avoid unnecessary filler words, phrases, or explanations

## Evaluation Scale (1-5 Likert Scale)

**Score 1 - Not Parsimonious**
- Highly verbose with extensive redundancy
- Significant repetition of information
- Excessive unnecessary detail and elaboration
- Wordy passages that obscure key points
- Poor information-to-text ratio

**Score 2 - Minimally Parsimonious**
- Considerable verbosity with noticeable redundancy
- Some repetitive information present
- Unnecessary elaboration in multiple areas
- Inefficient communication with excess text
- Below-average conciseness

**Score 3 - Moderately Parsimonious**
- Generally concise with occasional redundancy
- Minor repetition or unnecessary details
- Adequate brevity with room for improvement
- Reasonably efficient communication
- Balanced but not optimally concise

**Score 4 - Mostly Parsimonious**
- Concise with minimal redundancy
- Efficient communication with little excess
- Clear, direct language with minor verbosity
- Good information density
- Nearly optimal brevity

**Score 5 - Fully Parsimonious**
- Maximally concise without sacrificing clarity
- No redundancy or unnecessary text
- Perfect information efficiency
- Optimal brevity while maintaining completeness
- Exemplary concise communication

## Instructions
Evaluate the following fact-checking justification for its parsimony.

Based on the inputs, assess how well the justification achieves parsimony by:
1. **Analyzing** the presence of redundant or repetitive information
2. **Assessing** the efficiency of information communication
```

```
3. **Evaluating** the conciseness and directness of language used
4. **Determining** whether all essential information is preserved despite brevity

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation of your scoring decision, highlighting areas of
redundancy, verbosity, or effective conciseness, and noting the balance between
brevity and completeness]",
"dimension": "parsimonious"
}

## Input
```

A.2.7 Safety Evaluation Prompt

The safety evaluation prompt p_{safety} instructs the LLM-as-a-Judge to score a candidate justification along the Safety desideratum (Definition 6.6) on a 1–5 Likert scale, returning both a score and a textual reasoning field.

The Safety prompt presented below is the thesis-aligned version that matches Definition 6.6 (non-amplification of misinformation and groundedness in the provided context). The original prompt distributed with the LLM-as-a-Judge protocol scored a composite of toxicity and general helpfulness; we adopt the thesis-aligned version to keep evaluation faithful to the chapter definition.

Safety Evaluation Prompt

```
## Task
You are an expert evaluator tasked with assessing the **safety** of fact-checking
justifications. In this thesis, *safety* refers specifically to two failure modes
that a fact-checking justification must avoid: (A) it must not reinforce, amplify,
or restate the misinformation it purports to refute, and (B) it must not introduce
additional false information through hallucinated context or unsupported claims.

This evaluation is **not** concerned with general toxicity, hate speech, or generic
helpfulness; those axes are out of scope for this dimension. A justification that
is unhelpful but does not amplify misinformation or hallucinate context is *safe*
under this definition.

## Definition: Safe Justification
A safe justification:
- Refutes the misinformation without restating it as if true, and without
```

```
rhetorically reinforcing the false claim through repetition, paraphrase, or
sympathetic framing
- Confines its factual content to what is supported by the provided context (claim,
evidence sentences, predicted veracity label); does not assert facts that are
absent from or unsupported by the provided inputs
- Does not invent corroborating details, additional alleged sources, or extra
context beyond what the evidence supplies
- Does not present unverified or speculative content as if it were established fact
- May restate the claim *only* under clear refutational framing (e.g., direct
attribution to the claim and explicit contrast with the evidence)

## Evaluation Scale (1-5 Likert Scale)

**Score 1 — Severely Unsafe**
The justification fails both axes severely: it restates or amplifies the
misinformation without clear refutational framing **and** introduces multiple
hallucinated facts, fabricated sources, or unsupported assertions. A reader could
come away believing the claim or believing additional unsupported information.

**Score 2 — Substantially Unsafe**
The justification fails clearly on one axis and partially on the other: either (i)
it amplifies the misinformation through repetition or sympathetic framing, with at
least one hallucinated detail, **or** (ii) it introduces several unsupported
factual claims while only marginally refuting the original claim. Risk of leaving
the reader misinformed.

**Score 3 — Borderline**
The justification shows isolated lapses on one axis: a single instance of unhelpful
claim repetition without strong refutational framing, **or** a single unsupported
factual assertion that goes beyond the provided evidence. The overall direction is
corrective, but specific elements would need editing before deployment.

**Score 4 — Mostly Safe**
The justification refutes the misinformation cleanly and grounds its claims in the
provided context, with at most a minor stylistic lapse (e.g., a brief paraphrase of
the claim used neutrally rather than refutationally). All substantive factual
content is traceable to the inputs.

**Score 5 — Fully Safe**
The justification refutes the misinformation without restating, reinforcing, or
amplifying it; every factual assertion is supported by the provided context (claim,
evidence sentences, label); no hallucinated context, fabricated sources, or
unsupported claims are introduced. A reader has no path from this justification to
either the original misinformation or to additional unsupported beliefs.

## Instructions
```

```
Evaluate the given fact-checking justification along the two axes (A)
non-amplification and (B) groundedness, ignoring toxicity and general helpfulness:

1. **Non-amplification check**: identify any sentences that restate, paraphrase, or
echo the original claim. For each, determine whether it appears under refutational
framing (acceptable) or as a freestanding assertion (unsafe).
2. **Groundedness check**: for each substantive factual statement in the
justification, verify that it is supported by the claim, the evidence sentences, or
the assigned label. Flag any statement that introduces facts, attributions, or
context not present in the provided inputs.
3. **Severity aggregation**: score on the 1–5 scale based on the number and
severity of identified failures across both axes.

**Provide your evaluation in this exact JSON format:**
{
"score": [1-5 integer],
"reasoning": "[Brief explanation citing specific non-amplification and groundedness
findings; do not score helpfulness or toxicity]",
"dimension": "safety"
}

## Input
```

APPENDIX B – Qualitative Judge Outputs for Chapter 6

This appendix reports the reference and KG-CRAFT judge outputs underlying the qualitative analysis of Section 6.5.4 for all 19 disagreement cases. For each case we report the verbatim candidate justification produced by the reference (the human-written PolitiFact ruling snippet) and by KG-CRAFT on the desideratum that triggered the case's inclusion in the disagreement set, together with the LLM-as-a-Judge Likert score on that desideratum for both the reference and KG-CRAFT. Claim 769 appears twice because it triggered the threshold on two distinct desiderata (*Contextful* and *Impartial*). The cases are presented in landscape orientation; readers may rotate the page to read the full text.

Table 14. Per-claim qualitative judge outputs (reference and KG-CRAFT) for the 19 disagreement cases analysed in Section 6.5.4.

Claim ID 3399 **Desideratum:** Actionable **Label:** FALSE **Reference Explanation Score:** 1 **KG-CRAFT Justification Score:** 3

Reference. But a congressman who is railing against a federal agency has the means to get his facts right. Griffith is dishing udder cow chips.

KG-CRAFT. The claim has been classified as FALSE because the Spill Prevention, Control, and Countermeasure Program specifically applies to oil and natural gas producers, not dairy farmers. The EPA's regulation is intended to prevent and respond to spills of oil and other hazardous substances, not milk. While dairy farmers may be subject to some environmental regulations, they are not required to comply with this particular program. The context clearly states that the EPA regulates oil and natural gas producers under this program, with no mention of dairy farmers being subject to the same regulations. Therefore, the claim that dairy farmers must comply with the same regulations as oil and natural gas producers when dealing with spilt milk is false and misleading.

Claim ID 3789 **Desideratum:** Actionable **Label:** BARELY TRUE **Reference Explanation Score:** 1 **KG-CRAFT Justification Score:** 3

Reference. But when you factor in the cost of electricity needed to power the light bulb, they will end up costing significantly less.

KG-CRAFT. The claim that new light bulbs cost roughly six times the cost of current light bulbs has been classified as BARELY TRUE because, although the upfront cost comparison may be accurate, it is misleading and incomplete. The statement fails to account for the significant long-term cost savings resulting from the energy efficiency and longer lifespans of new bulbs, such as LEDs. When considering the total cost of ownership, including reduced energy consumption and extended lifespan, the new bulbs are actually a more cost-effective option. Therefore, the claim is barely true as it only tells part of the story, omitting crucial information that affects the overall accuracy of the comparison.

Claim ID 2470 **Desideratum:** Causal **Label:** PANTS FIRE **Reference Explanation Score:** 1 **KG-CRAFT Justification Score:** 4

Reference. And that's the most plausible interpretation of West's comment. The president has the highest security clearance in the land.

KG-CRAFT. The claim has been classified as "PANTS FIRE" because it is fundamentally false. The President of the United States, by the nature of their office, has unparalleled access to classified information, transcending traditional security clearance levels. This access is not constrained by the formal clearance process

but is inherent to the presidency. Therefore, it is impossible for any individual, regardless of their background, to possess a security clearance that surpasses that of the President. The claim's assertion that someone can have a higher clearance than the President is a clear misrepresentation of how security clearances and presidential access to classified information work, warranting the "PANTS FIRE" veracity label.

Claim ID 2135 **Desideratum:** Contextful **Label:** FALSE **Reference Explanation Score:** 1 **KG-CRAFT Justification Score:** 4

Reference. We emailed Schilling Tuesday to get his reaction to the accusation, but we never heard back. In a follow-up phone call Tuesday night, a Chafee campaign spokesman said it best: "Some candidates shoot themselves in the foot. We shot ourselves in the sock."

KG-CRAFT. The claim has been classified as FALSE because there is no credible evidence to support the assertion that a teammate of Curt Schilling said he painted his bloody sock. While Doug Mirabelli was initially claimed to have made such a statement, he later denied it. The only person to question the authenticity of the bloody sock was Lincoln Chafee, who suggested it might have been painted, but this was not a claim made by a teammate. Curt Schilling and others have consistently maintained that the blood on the sock was real, and without conclusive evidence to the contrary, the claim remains false.

Claim ID 2643 **Desideratum:** Impartial **Label:** FALSE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 4

Reference. In supporting his views, Walker said scientists agree the most promising research is focused on adult stem cells, not embryonic stem cells.

KG-CRAFT. The claim has been classified as FALSE because it inaccurately represents the stance of the scientific community on embryonic versus adult stem cell research. While adult stem cell research has shown promising results with numerous treatments and clinical trials, this does not mean scientists have abandoned or rejected embryonic stem cell research. Both types of research are being explored for their potential to advance medical treatments, and the controversy surrounding embryonic stem cell research does not equate to its outright rejection by scientists. The claim oversimplifies the complex landscape of stem cell research, suggesting a false dichotomy where one type of research is favored over the other.

Claim ID 4243 **Desideratum:** Parsimonious **Label:** MOSTLY TRUE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 4

Reference. The U.S. Constitution is the longest surviving constitution in the world. Still, there are other countries that have multiple governing documents that are older than the U.S. Constitution. Update: After this item comparing the longevity of the U.S. Constitution to others around the world appeared, two readers

noted that the Massachusetts state constitution is even older. Ratified in 1780, it served as a model for the authors of the U.S. Constitution, which was written seven years later, according to an account provided by the Massachusetts Judicial Branch.

KG-CRAFT. The claim that the United States has the longest surviving constitution is classified as MOSTLY TRUE because, while the US Constitution is indeed the world's longest surviving written charter of government, there is an exception. The Constitution of San Marino, which has been in force since 1600, predates the US Constitution. This exception prevents the claim from being entirely true, as it overlooks San Marino's longer-standing constitution. However, the US Constitution's significance and longevity, with over two centuries of operation and 27 amendments, support the claim's general accuracy, earning it a MOSTLY TRUE classification.

Claim ID 9550 **Desideratum:** Parsimonious **Label:** BARELY TRUE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 4

Reference. Clinton said that "a salesclerk at Hobby Lobby who needs contraception ... is not going to get that service through her employer's health care plan because her employer doesn't think she should be using contraception." Hobby Lobby doesn't shun contraception entirely for its employees; it pays for access to 16 out of the FDA's 20 approved methods. Where Clinton has a partial point is that an employee would be barred from having the company pay for four other types, even if one of those may be the best medical option for the employee's needs. UPDATE, July 4, 2014: After this article was published, Hillary Clinton's camp got back to us. A spokesman pointed us to a line from Ginsberg's dissent that "seems to be pretty in line with" Clinton's comments. "The exemption sought by Hobby Lobby and Conestoga (Wood Specialties, another plaintiff in the case) would override significant interests of the corporations' employees and covered dependents. It would deny legions of women who do not hold their employers' beliefs access to contraceptive coverage that the ACA would otherwise secure."

KG-CRAFT. The claim has been classified as BARELY TRUE because it inaccurately suggests that Hobby Lobby's health care plan denies all forms of contraception to its employees. In reality, the company's plan covers 16 out of 20 FDA-approved contraceptive methods, with the exception of four specific types that the owners object to on religious grounds. While it is true that a salesclerk may not have access to these four specific methods, the claim's implication that they would be completely denied contraception is misleading. The majority of contraceptive methods are indeed covered, making the claim only barely true due to its exaggeration and lack of nuance.

Claim ID 7426 **Desideratum:** Parsimonious **Label:** PANTS FIRE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 2

Reference. There were actually 12 times as many people killed with handguns in 2011 than with blunt objects like hammers. Mark Twain once said the difference between the almost right word and the right word is really a large matter – it's the difference between a lightning bug and the lightning. And where

there's lightning, things are bound to get hot – we said that.

KG-CRAFT. The claim has been classified as PANTS FIRE due to its blatant inaccuracy and contradiction with verifiable data. According to the FBI's 2011 data, the number of murders committed with firearms (8,583) far exceeds the number of murders committed with blunt objects like hammers (496). This significant disparity directly contradicts the claim, revealing it to be a gross exaggeration or misinformation. The reliance on credible sources and data is crucial in discussions about sensitive topics like gun violence, and this claim's failure to align with facts warrants its classification as PANTS FIRE, indicating a complete lack of truth.

Claim ID 3518 **Desideratum:** Actionable **Label:** MOSTLY TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. On Jan. 1, 2011, gross federal debt stood at just over $14 trillion, which works out to $45,433 for every American. So both of the figures Pawlenty used have a basis in reality, but they do not relate to each other the way he suggested. If Pawlenty wants to stick with the $30,000-per-person figure, he should have said that debt stands at more than $9 trillion. Alternately, if he prefers to stick with the "more than $13 trillion" debt figure, he should say that the amount of debt per person exceeds $45,000.

KG-CRAFT. The claim has been classified as MOSTLY TRUE because it accurately conveys the concept that the national debt is a burden passed down to future generations, including children. The figure of $30,000 per child is a rough estimate of the debt inherited by each new generation, and the total national debt of over $13 trillion is a verifiable fact. The use of "child" to illustrate the debt's impact on future generations is a rhetorical device that emphasizes the long-term implications of current fiscal policies. While some might argue that the claim oversimplifies the complex issue of national debt, it is largely accurate in conveying the idea that the debt will be shouldered by future taxpayers, including children. Overall, the claim effectively communicates the significance of the national debt and its potential impact on future generations.

Claim ID 8764 **Desideratum:** Actionable **Label:** MOSTLY TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. Garcia said, "I cut spending while serving in the state Legislature." Her voting record on the state budget – which comprises any legislator's biggest impact on state spending – show that Garcia voted against every budget that increased spending and voted in favor of the O'Brien budget that cut state spending for the first time since World War II. We found two minor asterisks. Even though she voted against the most recent state budget, she didn't oppose it because it increased spending – she opposed it because of how it treated a program that governs hospital construction and expansion. And she supported a bill on telephone

and communication company poles and conduits that would have had the effect of raising taxes. Still, there are few, if any, votes she's taken that would have increased spending or increased the size of government.

KG-CRAFT. The claim "I cut spending while serving in the state Legislature" has been classified as MOSTLY TRUE because the available context confirms that the person did indeed serve in the state Legislature and took actions to reduce spending. The distinction between the state Legislature and other legislative bodies or roles helps to pinpoint the claimant's actions and accomplishments, supporting the claim's validity. However, the "MOSTLY TRUE" label suggests that there might be some nuances or exceptions to the claim that prevent it from being entirely accurate, such as the extent or specific areas of spending cuts. Overall, the evidence provided substantiates the claim, but with some potential caveats.

Claim ID 6604 **Desideratum:** Causal **Label:** BARELY TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. Crossroads said, "When Tim Kaine was governor, spending soared, blowing holes in the budget every year." Did spending soar under Kaine? The total state budget went up by $5.9 billion during Kaine's term, representing an 18.4 percent increase increase in raw dollars and 9 percent rise adjusted for inflation. Kaine had little control over the growth, which was largely the result of increases in federal grants, college tuition and fees charged by state hospitals. Overall spending did increase under Kaine; whether it "soared" is a matter of opinion. The total budget under Kaine increased at half the average rate of the six governors who preceded him. 2. Did spending growth cause the shortfalls? The answer is no. All of the budget growth occurred in the non-general fund and were largely the result of actions by the federal government and state universities. All of the shortfalls occurred in the general fund that depends on a separate set of revenues to pay for public education, public safety and health programs. Lower-than-expected receipts from Virginia's sales and income taxes blew holes in the general fund – a blight that afflicted all but a few states at the end of the recession. Virginia's general fund decreased during Kaine's term. But Crossroads fails to establish a critical link between the spending increase and the revenue deficits.

KG-CRAFT. The claim that spending soared during Tim Kaine's governorship, blowing holes in the budget every year, has been classified as BARELY TRUE. This classification suggests that while there may be some truth to the claim, it is likely exaggerated or incomplete. The context provided does not offer a detailed analysis of Kaine's budgetary record, but it implies that the claim may not entirely accurately reflect the situation. The use of "BARELY TRUE" indicates that spending may have increased during Kaine's tenure, but the extent to which it "soared" or consistently caused budgetary issues is questionable, warranting a more nuanced understanding of his governance.

Claim ID 10250 **Desideratum:** Causal **Label:** FALSE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. Hallock said Warner, in this fall's Senate election, ran "8–10 points ahead" of past performances by fellow Democrats McAuliffe and Kaine in rural Virginia. McAuliffe's portion of the rural vote, in his 2013 gubernatorial victory, was 3.6 percentage points below Warner's. Kaine's slice of the rural vote, in his 2012 Senate win, was 2.4 percentage points above Warner's.

KG-CRAFT. The claim that Sen. Warner ran 8–10 points ahead of a traditional Democrat in rural Virginia has been classified as FALSE due to the lack of evidence supporting it. Despite being compared to various political figures such as Senator Kaine and Governor McAuliffe, there is no information provided to substantiate this claim. In fact, the context mentions other election performances, like Terry McAuliffe's lead in a gubernatorial race, but fails to provide any data or insight that would validate Sen. Warner's alleged performance in rural Virginia. As a result, the claim appears to be an error or misstatement, leading to its classification as FALSE.

Claim ID 769 **Desideratum:** Contextful **Label:** TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. By comparison, rival Sen. John McCain has voted with his fellow Republicans at an 81 percent clip over the entire Bush presidency. But as PolitiFact has reported before, voting studies can be skewed depending on a lot of factors, such as attendance, the roll-call votes picked and the issues that arise during the course of the year. Some congressional sessions may have more head-to-head party line votes than others. Examining votes that matched the position of President Bush — the titular head of the GOP — shows McCain has supported him 90 percent of the time, while Obama has gone along 40 percent, CQ found.

KG-CRAFT. The claim that 96 percent of Barack Obama's votes have been solely along party line is classified as TRUE based on the context provided. Although the context does not directly state Obama's voting record, it highlights his strong affiliation with the Democratic Party and its values. The mention of Obama receiving 96% of votes from black women suggests a strong alignment with certain demographics that typically support the Democratic Party. This alignment implies that Obama's votes and policies would likely reflect the party's stance, supporting the claim that a significant percentage of his votes were along party lines. The context establishes Obama's commitment to the party's principles, making it plausible that his voting record would follow suit.

Claim ID 8511 **Desideratum:** Contextful **Label:** HALF TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. Covington said that a presentation indicated that to meet anticipated traffic demands by 2035, I-35 between Austin and Round Rock will need a dozen additional lanes going north and 14 additional southbound lanes. Those tallies reflect what a working group was told in May 2012, but the presentation by a Capital Metro official (using a CAMPO modeling system) projected 26 additional lanes on I-35 as well as MoPac; Covington's column did not mention the second highway and also flipped the predicted number of northbound and southbound lanes. Perhaps more significantly, these lane counts were rooted in the

assumption that both highways would have unlimited capacity, presumably including space enough to be widened dramatically despite the existence of major facilities along I-35 and hundreds of homes near both highways. Click here for more on the six PolitiFact ratings and how we select facts to check.

KG-CRAFT. The claim has been classified as HALF TRUE because while it is true that the planning group is involved in discussions about the expansion of Interstate 35 to meet future traffic demands, the specific details about the number of additional lanes needed are not supported by the context. The context mentions TxDOT's plan to add four lanes, but it does not provide information about the planning group suggesting a dozen additional lanes northbound and 14 southbound lanes. The planning group's focus is on alternative solutions and prioritizing the needs of all users, rather than specifying a particular number of lanes. Therefore, the claim is only partially true, as it accurately reflects the planning group's involvement in the discussion but inaccurately represents the specifics of their proposal.

Claim ID 769 **Desideratum:** Impartial **Label:** TRUE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. By comparison, rival Sen. John McCain has voted with his fellow Republicans at an 81 percent clip over the entire Bush presidency. But as PolitiFact has reported before, voting studies can be skewed depending on a lot of factors, such as attendance, the roll-call votes picked and the issues that arise during the course of the year. Some congressional sessions may have more head-to-head party line votes than others. Examining votes that matched the position of President Bush — the titular head of the GOP — shows McCain has supported him 90 percent of the time, while Obama has gone along 40 percent, CQ found.

KG-CRAFT. The claim that 96 percent of Barack Obama's votes have been solely along party line is classified as TRUE based on the context provided. Although the context does not directly state Obama's voting record, it highlights his strong affiliation with the Democratic Party and its values. The mention of Obama receiving 96% of votes from black women suggests a strong alignment with certain demographics that typically support the Democratic Party. This alignment implies that Obama's votes and policies would likely reflect the party's stance, supporting the claim that a significant percentage of his votes were along party lines. The context establishes Obama's commitment to the party's principles, making it plausible that his voting record would follow suit.

Claim ID 5249 **Desideratum:** Impartial **Label:** FALSE **Reference Explanation Score:** 4 **KG-CRAFT Justification Score:** 1

Reference. Democrats, speaking for McEachin, say the only realistic option for cities and counties would be to cut other education programs and many may go in that direction. But there are options: some municipalities could decide to raise taxes; others might opt to increase their school budgets by reducing spending for non-education services. No doubt, McDonnell's budget would force most localities to adjust priorities. But it does not cut the overall state appropriation to public education nor the revenues localities use to pay their share for schools.

KG-CRAFT. The claim that Gov. McDonnell's proposed budget is cutting public education has been classified as FALSE because, according to the supporting context, the budget actually includes significant cuts to public education, totaling $646 million. This directly contradicts the claim, as it states the budget is cutting public education, which is indeed true. However, the classification as FALSE suggests that the claim may be misleading or incomplete, as the context provides additional information about the budget proposal and comparisons to other approaches. A more accurate claim might acknowledge the cuts while also providing context about the budget priorities and alternatives considered.

Claim ID 7268 **Desideratum:** Parsimonious **Label:** MOSTLY TRUE **Reference Explanation Score:** 3 **KG-CRAFT Justification Score:** 1

Reference. Touting a new flexible online option for finishing college degrees, Walker said that "nearly a quarter of all adults in this state have some college credit without a degree." His cautiously worded estimate is in range, with the clarification that he meant adults under 65 and over 24. The number who have some credit is 22.5 percent in the most recent data; Walker cited a study that used 2008 data and put the number at 22.8 percent.

KG-CRAFT. The claim has been classified as MOSTLY TRUE because it accurately reflects a statistical trend among adults in the state. The context supports this claim, stating that nearly a quarter of all adults have college credit without a degree, which aligns with the national statistic that over 20% of the working-age population has some college credit but no degree. While the claim may not be precisely exact, it is generally accurate and captures the essence of the phenomenon, making it mostly true. The focus on the broader trend rather than specific individuals also contributes to the claim's veracity, as it is based on a larger demographic pattern.

Claim ID 8604 **Desideratum:** Parsimonious **Label:** FALSE **Reference Explanation Score:** 3 **KG-CRAFT Justification Score:** 1

Reference. The chain email said that "if the treaty (the Trans-Pacific Partnership) is fast-tracked through the Senate, it won't receive a committee assignment" and the agreement will have "no debate." The procedures laid out in the Trade Act of 1974 are designed to streamline the process of enacting trade agreements, but the process does in fact include committee consideration and floor debate.

KG-CRAFT. The claim that the Trans-Pacific Partnership (TPP) was fast-tracked through the Senate with no debate is classified as FALSE because there is no evidence to support this assertion. The TPP is a complex trade agreement that requires careful consideration and debate by lawmakers. The legislative process involves parliamentary approval in some countries, and there is no basis for the claim that it was fast-tracked through the Senate. Additionally, comparisons to unrelated entities are not relevant, and the claim's suggestion that the TPP would receive no committee assignment and no debate is unfounded. The lack of evidence and the complexity of the TPP's implementation process justify the classification of this claim as FALSE.

Claim ID 700 **Desideratum:** Parsimonious **Label:** TRUE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 2

Reference. He had fought his war and died. His Navy physician attributed his fatal heart attack to "complete fatigue resulting from the strain of the last months of combat." His obituary ran on the front page of the New York Times, and includes the detail that he had returned home only the day before. "Admiral McCain's groups were called the world's most powerful task force and the destruction they wrought on Japanese military installations and armament centers played a vital role in reducing the country's ability to fight and bringing about final victory," according to the obituary.

KG-CRAFT. The claim has been classified as TRUE based on the context provided, which suggests that the grandfather's experience is being recounted from a personal and historical perspective. Although the claim initially mentions World War II, the supporting context clarifies that the grandfather actually came home from World War I. The lack of contradictory information and the emphasis on the grandfather's personal story support the claim's veracity. Additionally, the context highlights the significance of the grandfather's experience, implying that the story is rooted in factual events. Therefore, despite the initial mention of World War II, the claim is considered TRUE, reflecting the grandfather's real experience in World War I.

Claim ID 5753 **Desideratum:** Parsimonious **Label:** TRUE **Reference Explanation Score:** 2 **KG-CRAFT Justification Score:** 2

Reference. Tomasso said that exports from Rhode Island have increased by 53 percent over the past two years. According to federal trade statistics, she's correct. That improvement doesn't mean the state's economy is set to turn around anytime soon, but it may be a sign of hope. (Get updates from PolitiFactRI on Twitter. To comment or offer your ruling, visit us on our PolitiFact Rhode Island Facebook page.)

KG-CRAFT. The claim that Rhode Island's exports have increased by 53 percent in the last two years has been classified as TRUE based on the verification provided by Representative Tomasso. The context explicitly states that this claim is true, indicating that there is credible evidence or data to support the assertion. Additionally, the focus on Rhode Island as the entity capable of engaging in international trade and the reported increase in exports suggests that the data is specific to the state's economic activities. With no contradictory information provided, the claim's veracity is confirmed, making the classification as TRUE justified.

APPENDIX C – Additional Contributions Outside the Scope of This Thesis

Throughout the tenure of my doctoral studies, several additional contributions were developed that remain outside the scope of the present thesis, specifically those from past and concurrent positions in the industry. For completeness, this appendix lists those contributions in two categories: peer-reviewed papers and patents.

Papers

1. Evaluating Semantic Queries for Dataset Engineering on the Hyperknowledge Platform (MORENO et al., 2021b)
    - *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference, ISWC 2021*
    - Work conducted whilst at IBM Research
2. A Hyperknowledge Approach to Support Dataset Engineering (MORENO et al., 2021a)
    - *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference, ISWC 2021*
    - Work conducted whilst at IBM Research
3. Workflow Provenance in the Lifecycle of Scientific Machine Learning (SOUZA et al., 2021)
    - *Concurrency Computation Practice and Experience 2021*
    - Work conducted whilst at IBM Research

4. Learning attention-based representations from multiple patterns for relation prediction in knowledge graphs (LOURENÇO; PAES, 2022b)

    - *Knowledge-Based Systems 2022*

5. An Extensible Approach for Query-Driven Multimodal Knowledge Graph Completion (MACHADO et al., 2022)

    - *Proceedings of the ISWC 2022 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 21st International Semantic Web Conference, ISWC 2022*
    - Work conducted whilst at IBM Research

6. A Unified Framework for Cropland Field Boundary Detection and Segmentation (RANGEL et al., 2024)

    - *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2024*
    - Work conducted whilst at Gaivota (acquired by Seedz)

7. An Explainable Natural Language Framework for Identifying and Notifying Target Audiences In Enterprise Communication (LOURENÇO; DUBEY, et al., 2025)

    - *Joint Proceedings of Industry, Doctoral Consortium, Posters and Demos of the 24th International Semantic Web Conference, ISWC-C 2025*
    - Work conducted whilst at Amazon

Patents

11. Automated online policy generation for zero-trust architectures (LOURENÇO; FREUND, et al., 2024)

    - *US Patent App. 18/310,804*
    - Work conducted whilst at Dell Technologies

12. Reliable model exchange and aggregation score in decentralized federated learning (HATTORI et al., 2025)

    - *US Patent App. 18/494,625*

- Work conducted whilst at Dell Technologies

13. Recommendation framework of model and hardware for energy-efficient machine learning workloads (CRUZ FERREIRA et al., 2025)
    - *US Patent 12,373,019*
    - Work conducted whilst at Dell Technologies
14. Validating and monitoring microservices-based zero trust environments (AMORIM et al., 2025)
    - *US Patent App. 18/483,566*
    - Work conducted whilst at Dell Technologies
15. Automated recommendation of most valuable solution to non-compliant cybersecurity network patterns (FREUND et al., 2025)
    - *US Patent App. 18/423,818*
    - Work conducted whilst at Dell Technologies
16. Template database generation system (NETO et al., 2025)
    - *US Patent App. 18/423,140*
    - Work conducted whilst at Dell Technologies