Advancing Explainable AI: A Global Context-Aware Evidence Retrieval Framework for Interpretable Fact Verification

Manju Vallayil Vijayalekshmi

A thesis submitted to Auckland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

2025

School of Engineering, Computer & Mathematical Sciences

To my daughter, **Sarangi**, whose wit, wisdom, and unfiltered honesty kept me grounded through this journey. Your sharp humor reminded me to laugh at the struggles, your curiosity pushed me to keep learning, and your love gave me the strength to move forward. May you always question boldly, dream fearlessly, stay compassionate, and let your unique spark shine in everything you do.

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Manju Vallayil Vijayalekshmi
February 2025

Co-authorship Contribution

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name: Manju Vallayil Vijayalekshmi

Signature: **Date:** 19 Feb 2025

Supervisor Name: Dr. Parma Nand

Signature: **Date:** 19 Feb 2025

Chapter Number:	3
Manuscript Title:	Explainability of Automated Fact Verification Systems: A Comprehensive Review
Publication Status:	Published
Reference if Published:	Vallayil, M., Nand, P., Yan, W. Q., & Allende-Cid, H. (2023). *Explainability of Automated Fact Verification Systems: A Comprehensive Review*. *Applied Sciences*, *13*(23), 12608. https://doi.org/10.3390/app132312608
AUTHOR SURNAME: *(order as per manuscript)*	**CONTRIBUTION** *(as listed below)*
M. Vallayil	– Conception, Methodology, Formal-analysis, Writing, Visualization
P. Nand	– Supervision, Conception, Review and editing
W. Q. Yan	– Supervision, Review and editing
H. Allende-Cid	– Review and editing

Chapter Number:	5
Manuscript Title:	Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems
Publication Status:	Published
Reference if Published:	Vallayil, M., Nand, P., & Yan, W. Q. (2024). *Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems*. In *ACIS 2024 Proceedings* (Paper No. 101).
AUTHOR SURNAME: *(order as per manuscript)*	**CONTRIBUTION** *(as listed below)*
M. Vallayil	– Conception, Methodology, Software, Writing, Visualization
P. Nand	– Supervision, Resources, Review and editing
W. Q. Yan	– Supervision, Review and editing

Chapter Number:	7
Manuscript Title:	CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI
Publication Status:	Published
Reference if Published:	Vallayil, M., Nand, P., Yan, W. Q., Allende-Cid, H., & Vamathevan, T. (2025). *CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI*. *Applied Sciences*, *15*(4), 1970. https://doi.org/10.3390/app15041970
AUTHOR SURNAME: *(order as per manuscript)*	**CONTRIBUTION** *(as listed below)*
M. Vallayil	– Conception, Methodology, Software, Writing, Visualization, Data Analysis
P. Nand	– Supervision, Resources, Review and editing
W. Q. Yan	– Supervision, Review and editing
H. Allende-Cid	– Review and editing
T. Vamathevan	– Review and editing

Chapter Number:	9
Manuscript Title:	Unsupervised Thematic Context Discovery for Explainable AI in Fact Verification: Advancing the CARAG Framework
Publication Status:	Accepted at KDIR 2025
Reference if Published:	–
AUTHOR SURNAME: *(order as per manuscript)*	**CONTRIBUTION** *(as listed below)*
M. Vallayil	– Conception, Methodology, Software, Writing, Visualization
P. Nand	– Supervision, Resources, Review and editing
W. Q. Yan	– Supervision, Review and editing
H. Allende-Cid	– Review and editing

Acknowledgements

I would like to express my deepest gratitude to my primary supervisor, Dr. Parma Nand, whose guidance and support have been instrumental in shaping this research. His ability to provide the right balance of independence and direction allowed me the freedom to explore innovative ideas while ensuring I stayed on the right path. His encouragement to think beyond conventional boundaries has been invaluable. I am also deeply grateful to Associate Professor Wei Qi Yan, whose unwavering commitment to publication excellence and meticulous attention to research standards significantly contributed to the refinement of this work. His prompt feedback, availability even during weekends and holidays, and keen focus on high-quality dissemination have been truly inspiring. My sincere thanks also go to Associate Professor Héctor Allende-Cid, whose deep expertise and critical insights have strengthened this research immensely. His thoughtful comments and constructive critiques have helped sharpen the clarity and rigour of my work. I also extend my deepest gratitude to my mentor and sister-like guide, Dr. Thamilini, whose guidance and wisdom have been a constant source of strength and encouragement throughout this journey.

I would like to express my sincere gratitude to the School of Engineering, Computer and Mathematical Sciences(ECMS) and Te Ipukarea Research Institute for their support throughout my research journey. In particular, I am deeply thankful to Professor Tania Ka'ai and Tania Smith-Henderson for their invaluable guidance and encouragement. I also extend my appreciation for the financial support provided, which enabled me to focus on my research and contribute meaningfully to this work. I would also like to extend my gratitude to Bumjun Kim, for his assistance and support throughout my research. His technical expertise and prompt guidance were invaluable in overcoming challenges during this work.

Finally, I extend my heartfelt gratitude to the 'Cosmic Grace' that has illuminated my path, embodied in the unwavering support of my parents, Manoharan and Vijayalekshmi, and my brother, Manu. A special note of appreciation goes to my husband, Harikumar, whose unwavering support and encouragement have been my greatest source of strength throughout this journey. May everyone be fortunate enough to have a life partner who celebrates their victories as their own.

Abstract

The increasing reliance on Artificial Intelligence (AI) for decision-making across various domains necessitates the development of explainable AI (XAI) systems that enhance transparency and interpretability. However, despite growing research interest in XAI, ensuring transparency in Automated Fact Verification (AFV) remains a significant challenge, largely due to the architectural complexity of these systems and their reliance on local post-hoc explanations, which focus on claim-level context or annotated evidence while often overlooking broader thematic connections. Meanwhile, retrieval-based approaches like Retrieval-Augmented Generation (RAG) integrate external information but may retrieve loosely related or overly broad evidence, affecting the coherence and factual accuracy of the generated explanations. This thesis addresses this gap by introducing CARAG, a Context-Aware Retrieval and Explanation Generation framework that integrates both local and global perspectives to improve retrieval transparency and explanation coherence.

The research unfolds through four key manuscripts: (1) a comprehensive literature review that identifies major explainability gaps in AFV, particularly the lack of explanation-focused datasets and the overemphasis on local transparency, (2) an exploration of thematic context discovery as a means to uncover a claim's broader, non-local context within the fact-checking corpus, laying the groundwork for context-aware evidence retrieval, (3) the introduction of CARAG, which builds on thematic context discovery to enhance evidence selection and explanation generation in AFV pipelines, alongside the development of FactVer, a dataset designed to support explainability-driven fact verification research, and (4) CARAG-u, an unsupervised extension that eliminates dependency on predefined thematic labels, making the CARAG framework more adaptable across diverse verification settings.

Empirical evaluations demonstrate that integrating thematic context into retrieval enhances AFV explainability, bridging the gap between claim-specific justifications and broader knowledge patterns within fact-checking corpora. The findings contribute to the development of scalable and interpretable AFV models, reinforcing trust and transparency in AI-driven fact verification.

Publications

List of Published/Submitted Research Articles

#	Publications*	Status
1	Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. (2023). Explainability of Automated Fact Verification Systems: A Comprehensive Review. *Applied Sciences*, 13(23), 12608. DOI 10.3390/app132312608. *Contribution: Manju Vallayil (85%), Parma Nand (10%), Wei Qi Yan (3%), Héctor Allende-Cid (2%)*	Published
2	Manju Vallayil, Parma Nand, and Wei Qi Yan. (2024). Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems. *35th Australasian Conference on Information Systems (ACIS) 2024 Proceedings*, number 101. *Contribution: Manju Vallayil (85%), Parma Nand (11%), Wei Qi Yan (4%)*	Published
3	Manju Vallayil, Parma Nand, Wei Qi Yan, Héctor Allende-Cid, and Thamilini Vamathevan. (2025). CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI. *Applied Sciences*, 15(4), 1970. DOI 10.3390/app15041970. *Contribution: Manju Vallayil (85%), Parma Nand (7%), Wei Qi Yan (3%), Héctor Allende-Cid (3%), Thamilini Vamathevan (2%)*	Published
4	Manju Vallayil, Parma Nand, Wei Qi Yan, and Héctor Allende-Cid. (2025). Unsupervised Thematic Context Discovery for Explainable AI in Fact Verification: Advancing the CARAG Framework. *Contribution: Manju Vallayil (85%), Parma Nand (10%), Wei Qi Yan (3%), Héctor Allende-Cid (2%)*	Accepted at KDIR 2025
5	Manju Vallayil and Parma Nand. (2025). *factver_master (Revision 0ff0df9)*. Hugging Face. doi: 10.57967/hf/5772. *Contribution: Manju Vallayil (50%), Parma Nand (50%)*	Dataset released on Hugging Face

** All journals and conferences listed above are Scopus-Indexed and Peer-Reviewed*

Table of contents

List of figures

List of tables

Chapter 1

Introduction

In an era of abundant yet often unreliable information, the need for transparency and interpretability in AI-driven decision-making has never been more crucial. This is especially critical in high-stakes domains such as journalism, policy-making, legal decision-making, and Automated Fact Verification (AFV), where AI-generated decisions directly impact public trust. The risks of opacity in such systems have become increasingly evident in recent years through real-world crises, such as the spread of COVID-19 misinformation, which influenced public health responses, and political misinformation campaigns like the Cambridge Analytica data scandal, which shaped electoral discourse. In such cases, the lack of transparent claim verification has led to significant societal repercussions, reinforcing the urgency of explainability. Similarly, as AI systems continue to influence decision-making at scale, this growing reliance has intensified concerns about their interpretability and trustworthiness. To address these challenges, Explainable Artificial Intelligence (XAI) has emerged as a key research area, aiming to enhance transparency across diverse AI applications.

> *"The stated goal of explainable artificial intelligence (XAI) was to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems."* DARPA's Explainable AI (XAI) Program: A Retrospective (Gunning et al., 2021, p. 1).

However, while XAI has advanced significantly in recent years, its application in AFV remains largely underexplored (Kotonya and Toni, 2020a). Nevertheless, transparency is crucial for fostering trust in fact verification, where decisions must be both

evidence-driven and interpretable. This gap may stem from the inherent complexity of AFV architectures, which involve multiple interconnected stages of retrieval and reasoning. AFV systems traditionally followed a three-stage pipeline: document retrieval, evidence selection, and Recognizing Textual Entailment (RTE) (Chen et al., 2022b; DeHaven and Scott, 2023; Jiang et al., 2021; Krishna et al., 2022; Soleimani et al., 2019; Thorne et al., 2018b; Zhong et al., 2020). Early AFV approaches verified claims using either pre-annotated evidence or dynamically retrieved documents, as seen in the FEVER benchmark (Thorne et al., 2018b), where claims are fact-checked against Wikipedia. The introduction of transformer architectures (Vaswani et al., 2017) brought significant improvements to the AFV pipeline, enhancing contextual understanding and retrieval efficiency. However, these advancements came at the cost of explainability, further increasing the opacity of AFV models and making it difficult to trace how claims are verified. In response, post-hoc explanation techniques (Shu et al., 2019) have emerged in XAI, aiming to provide justifications for model decisions rather than making the underlying reasoning process inherently transparent. In AFV, these methods analyze model outputs to offer insights into fact verification predictions. However, as AFV systems further evolved, the increasing reliance on retrieval-driven verification introduced additional challenges in ensuring explainability. A major breakthrough in AFV came with the integration of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and Large Language Models (LLMs) into the verification pipeline. While RAG dynamically retrieves relevant evidence at inference time and reduces the need for extensive fine-tuning, it also exacerbates the AI/ML Black Box Paradox (Ali et al., 2023; Gunning et al., 2021), making it difficult to trace how and why certain evidence pieces are selected and how they influence final veracity assessments.

Further adding to this challenge, most existing XAI approaches in AFV focus on local explainability (Kotonya and Toni, 2020a), justifying individual claim verifications while overlooking the broader thematic context that shapes fact-checking outcomes. While local transparency helps interpret specific predictions, it fails to capture how a claim relates to the broader knowledge base, where context can provide additional insights into its veracity. A claim's alignment with overarching thematic patterns across the dataset can refine retrieval decisions and influence the quality of generated explanations. Thus, advancing AFV explainability requires integrating both local and global perspectives, ensuring that verification incorporates not only claim-specific reasoning but also its contextual relevance within the larger fact-checking corpus.

In addition to challenges in retrieval transparency and the limitations of local explainability, studies exploring XAI-driven AFV (Kotonya and Toni, 2020a) highlight a key limitation: the lack of explanation-focused datasets. Existing benchmark datasets like FEVER (Thorne et al., 2018a) and MultiFC (Augenstein et al., 2019), while widely used in AFV research, are primarily designed for claim verification rather than explanation generation. As Stammbach and Ash (2020) highlighted, datasets that facilitate both claim verification and meaningful explanation learning are essential for advancing explainability in AFV. Without such datasets, models struggle to effectively generate interpretable justifications, further constraining the development of XAI-driven fact verification approaches.

To consolidate the key challenges outlined so far, this thesis identifies three fundamental gaps in XAI-driven AFV that hinder transparency, interpretability, and systematic evaluation:

- **Opaque Evidence Retrieval and Its Impact on Post-Hoc Explanations:** Retrieval-driven verification and explanation generation, particularly with RAG and LLMs, make evidence selection difficult to interpret, which in turn affects the reliability and coherence of AI-generated post-hoc explanations generated as part of the evidence-based claim verification process.
- **Limited Scope of Explainability:** Existing methods focus on local explainability, overlooking how broader thematic context influences a claim's veracity assessment within the fact verification process.
- **Lack of Standardized Evaluation for XAI in AFV:** Current AFV datasets prioritize claim verification but lack structured benchmarks designed to systematically evaluate explanation quality and thematic alignment in XAI-driven fact verification.

To address these challenges, this thesis introduces CARAG (Context-Aware Retrieval and Explanation Generation) (Vallayil et al., 2025), an innovative evidence retrieval mechanism that enhances AFV by tackling retrieval opacity and the limitations of local explainability. As discussed earlier, while conventional query-based retrieval methods such as RAG reduce the need for extensive fine-tuning and improve retrieval efficiency, they also increase system opacity, making it difficult to trace how evidence is selected. CARAG retains the flexibility of RAG while enhancing interpretability by introducing a novel approach to enriching query representations with thematic embeddings. This ensures that retrieved documents are not only relevant

but also thematically aligned with the claim, improving retrieval transparency and strengthening the coherence of AI-generated post-hoc explanations. In the upcoming chapters, this thesis further explores how thematic embeddings are derived from a subset of the fact verification corpus and integrated into query representations to enable a more context-aware retrieval process.

Building on CARAG, this thesis later introduces CARAG-u, an unsupervised extension that eliminates reliance on predefined thematic annotations, making explainability more adaptable to broader fact verification contexts. While CARAG improves retrieval transparency through structured thematic embeddings, CARAG-u dynamically discovers thematic structures without requiring labeled data, extending its applicability beyond predefined datasets. By removing dependency on pre-annotated claim-evidence pairs, CARAG-u takes a step toward data-agnostic explainability in AFV. Empirical evaluations against standard RAG models demonstrate that CARAG-u maintains thematic coherence, improves explanation consistency, and enhances retrieval quality, offering a scalable approach to explainability in fact verification.

Furthermore, this research presents FactVer, a benchmark dataset designed to support both fact verification and explainability research in AFV. Unlike conventional datasets that primarily focus on claim verification, FactVer incorporates structured evidence relationships and human-generated explanations, enabling a more robust evaluation of retrieval transparency and explanation quality. By organizing claims and evidence across diverse thematic domains, such as climate change, COVID-19, and electric vehicles, FactVer facilitates the assessment of both local and global explainability in fact verification. Additionally, its structured annotations provide a standardized framework for analyzing thematic relevance and explanation fidelity, addressing the lack of datasets designed to support XAI research in AFV. Further details on FactVer's structure, annotation process, and its role in evaluating retrieval interpretability, explanation coherence, and thematic alignment, will be discussed in later chapters.

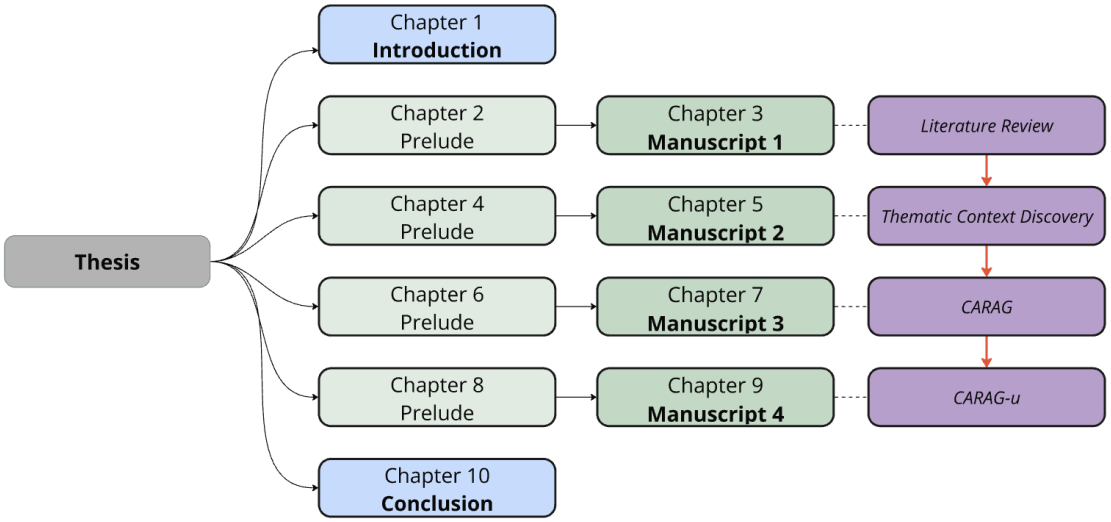


Fig. 1.1 The thesis follows the Manuscript Structure format, consisting of an introduction, four manuscripts, prelude sections linking each manuscript, and a concluding chapter. The research progression is depicted with a red arrow, tracing the development from Literature Review to Thematic Context Discovery, which establishes the foundation for CARAG, and further evolving into CARAG-u. This structured flow highlights how each stage builds upon the previous one, reinforcing the research progression and its evolving contributions to explainability in AFV.

1.1 Thesis Structure

This thesis follows the Manuscript Structure format (*Format Two*[1]), where the core research contributions are presented through a series of published and submitted manuscripts, supported by synthesis sections that ensure coherence in the overall narrative. As depicted in Figure 1.1, the research follows a structured progression, with a red arrow indicating the stepwise development from the initial Literature Review to Thematic Context Discovery, which lays the groundwork for CARAG and subsequently extends into CARAG-u.

The thesis begins with Chapter 1, which establishes the research context, problem statements, thesis structure, and contributions. Each manuscript is preceded by a prelude chapter that links it to the overarching research narrative. Chapter 3

[1]Refer to University Handbook 2025, https://www.aut.ac.nz/__data/assets/pdf_file/0003/796224/AUT-Postgraduate-Handbook-V1.0-January-2025-Final.pdf, pages 96–100.

presents a comprehensive literature review, identifying key explainability gaps in AFV. Chapter 5 investigates thematic context discovery as a foundational step for enhancing retrieval transparency. Chapter 7 builds upon this by integrating thematic context into the AFV pipeline through CARAG and introducing FactVer, the benchmark dataset designed to assess retrieval transparency and explanation quality. Chapter 9 further advances this work by extending CARAG into CARAG-u, an unsupervised variant that eliminates reliance on predefined thematic labels, making explainability more adaptable to fact verification beyond manually labeled datasets. Finally, Chapter 10 consolidates the research findings, highlights key contributions, and outlines future research directions.

1.2 Research Questions

Our work is structured around the following research questions:

- **RQ1:** How can we systematically assess the state of explainability in AFV systems and identify key limitations in existing methodologies, particularly in the context of local and global explainability?
- **RQ2:** How can we identify the non-local (thematic) context of a claim beyond its annotated evidence, and what methods can be used to extract and represent this broader contextual information?
- **RQ3:** How can integrating non-local (thematic) context improve explainability in AFV models?
- **RQ4:** Can AFV systems be enhanced with unsupervised explainability mechanisms that generalize across open-domain fact verification tasks without relying on pre-annotated datasets?

1.3 Research Methodology

This thesis adopts a structured research methodology following Creswell (2009), which defines research design through three key components: Philosophical worldview, Strategy of inquiry, and Research methods. Given the interdisciplinary nature of this research, spanning XAI and AFV, the methodology is designed to accommodate both theoretical exploration and empirical validation.

Philosophical Worldview: This research follows a pragmatic worldview (Creswell, 2009), prioritizing solutions to practical research problems over rigid epistemological stances. Pragmatism allows methodological flexibility, integrating both qualitative and quantitative approaches (Patton, 1990). Given the focus on identifying explainability gaps, developing methodologies (CARAG, CARAG-u), and empirically evaluating their effectiveness, this approach ensures both theoretical grounding and real-world applicability.

Strategy of Inquiry: This research employs a Sequential Mixed Methods approach (Creswell, 2009), integrating quantitative and qualitative methodologies for a comprehensive evaluation of the proposed frameworks. The study incorporates structured quantitative analysis to assess thematic alignment, retrieval accuracy, and explanation coherence across multiple cases, identifying overarching patterns in verification performance. Additionally, qualitative case studies provide an in-depth exploration of thematic influences on explanation generation. The sequential design ensures that quantitative findings establish broader trends, while qualitative assessments contextualize these insights.

Research Methods: The research methodology aligns with the pragmatic worldview and Sequential Mixed Methods strategy, integrating empirical evaluation with qualitative thematic analysis. The quantitative component involves statistical comparisons across multiple retrieval and verification models, evaluating thematic alignment and explanation coherence through structured metrics. The qualitative component comprises case studies examining thematic reasoning and interpretability within the CARAG and CARAG-u frameworks. This approach empirically validates model performance while contextualizing thematic influences on explainability, reinforcing interpretability and scalability in AFV.

Table 1.1 Overview of the Research Methodology Components

Component	Description
Philosophical Worldview	Pragmatic Worldview: Emphasizes practical problem-solving over rigid epistemological frameworks, integrating both qualitative and quantitative methods.
Strategy of Inquiry	Sequential Mixed Methods: Combines structured quantitative evaluation with qualitative case studies for a comprehensive assessment of explainability in AFV.
Research Design	Multi-Stage Empirical Approach: Progresses from thematic discovery and dataset construction to empirical validation, ensuring iterative refinement and practical applicability.

As summarized in Table 1.1, this research follows a pragmatic worldview and employs a Sequential Mixed Methods approach to investigate explainability in AFV, integrating empirical evaluation with qualitative case studies. Following Creswell (2009), the research design is structured to address the complexity of AFV explainability while also aligning with the interdisciplinary needs of AI and fact verification researchers. These considerations ensure the study effectively addresses real-world explainability challenges while contributing to XAI advancements.

1.4 Key Contributions

In summary, this thesis bridges critical gaps in explainability within AFV by introducing new methodologies, a benchmark dataset, and explainability-driven retrieval frameworks that enhance both local and global transparency. The research systematically assesses explainability limitations in AFV (addressing RQ1 in Chapter 3), and introduces thematic discovery to identify and represent broader claim context beyond annotated evidence (addressing RQ2 in Chapter 5). Through the development of FactVer, a dataset designed to support explainability-driven evaluation, and the introduction of CARAG, which integrates thematic retrieval mechanisms to improve context-aware evidence selection, this thesis advances AFV beyond traditional black-box models (addressing RQ3 in Chapter 7). Finally, CARAG-u

extends this framework by eliminating reliance on predefined thematic labels, enabling explainability in open-domain fact verification tasks through unsupervised thematic discovery (addressing RQ4 in Chapter 9). By addressing these core research questions, this work establishes a foundation for scalable, interpretable AFV systems that balance performance and transparency in real-world applications.

Chapter 2

Prelude - Manuscript 1

Manuscript 1 (Chapter 3) systematically examines explainability in AFV, analyzing how architectural, methodological, and dataset-related factors influence transparency in fact verification. This review identifies critical gaps, particularly the overreliance on local post-hoc explanations, and emphasizes the need for integrating both localized explanations for individual claims and broader thematic insights. Additionally, it highlights the absence of datasets designed to support explainability research. By bridging insights from XAI and AFV, Manuscript 1 establishes key research directions explored in subsequent studies, particularly in thematic context discovery (Manuscript 2), which lays the groundwork for later contributions, including CARAG and FactVer (Manuscript 3).

Chapter 3

Explainability of Automated Fact Verification Systems: A Comprehensive Review (Manuscript 1)

3.1 Introduction

Advances in Artificial Intelligence (AI), particularly transformer architecture (Vaswani et al., 2017), and the sustained success it has brought in transferring learning approaches in natural language processing, have led to advances in Automated Fact Verification (AFV). AFV systems are increasingly used in AI applications, making it imperative that AI assisted decisions are also accompanied by reasoning, especially in sensitive sectors like medicine and finance (Ali et al., 2023). In addition, researchers have also increasingly recognized the significance of AFV in the modern media landscape, where the rapid dissemination of information and misinformation has become a pressing concern (Guo et al., 2022). Consequently, AFV systems have become pivotal in addressing the challenges posed by the spread of online misinformation, particularly in verifying claims and assessing their accuracy based on evidence from textual sources (Du et al., 2022).An AFV pipeline involves the sub tasks for collecting evidence related to a claim, sorting most relevant evidence sentences, and predicting the veracity of the claim. Some systems such as Hassan et al. (2017) follow an additional step in the prelim to detect whether a claim is check-worthy or not before commencing on the other sub tasks in the pipeline. Besides these sub tasks, recent studies like Chen et al. (2022a) have started exploring how to generate automatic explanations as the reason for veracity prediction. However,

not as much effort has been put into the explanation functionality of AFV compared to the strong progress made over the past few years both in fact checking technology and data sets (Kotonya and Toni, 2020a). The lack of focus on explanation is behind the growing interest in explainable AI research (Kotonya and Toni, 2020a). Explainable AI[1] aims to provide the reasoning behind the decision (prediction) made, in contrast to the `black box' impression[2] of machine learning where even the AI practitioners fail to explain the reason behind a particular decision made by an AI system they designed. Similarly, the goal of explainable AFV systems is to go beyond simple fact verification by generating interpretations that are grounded in facts and that communicate in a way that is easily understood and accepted by humans. While there is broad agreement in the research community on the importance of the explainability of AI systems (Došilović et al., 2018; Guidotti et al., 2018; Kim, 2018), there is much less agreement on the current state of explainable AFV. The latest studies on the verification of facts (Das et al., 2023; Olivares et al., 2023; Rani et al., 2023) do not cohere around an aligned view on the subject. While researchers like Olivares et al. (2023) state that "Modern fact verification systems have distanced themselves from the black-box paradigm", Rani et al. (2023) contradict this by stating, modern AFV systems estimate the truthfulness "using numerical scores which are not human-interpretable". The same impression, as articulated in the latter statement, can be drawn from the literature review of state-of-the-art AFV systems. Another of the most recent arguments supporting this view is Das et al. (2023). They assert that, despite being a "nontrivial task", explainability of AFV is "mostly unexplored" and "needs to evolve" compared to the developments in explainable NLP.

This issue is further exacerbated by the fact that providing justifications for claim verdicts has always been a crucial aspect of human fact checking (Guo et al., 2022; Kotonya and Toni, 2020a). Therefore, it becomes evident that the transition from manual to automated fact checking falls short of achieving its intended human aspect to the functionality for `Automated Fact Verification' unless there is a clear incorporation of explainability.

In this paper, exploration of the concepts of explainability in XAI is initiated in Section 3.2, followed by a specific focus on its implementation in AFV in Section 3.3. By defining explainability within the context of AFV and introducing the architec-

[1](Also known as Interpretable AI or Explainable Machine Learning. Although used interchangeably the the subtle difference between explainability and interpretability, as is that the latter is not necessarily easily understood by those with little experience in the field unlike the former.

[2]https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

tural, methodological, and dataset-based aspects for discussing interpretations, the objective is to support and inspire research and implementations that can initiate the process of bridging the current explainability gap in AFV. The emphasis lies on the importance of datasets in achieving global explainability in AFV systems, suggesting that they should be a major focus of future research. Building upon these considerations, this paper has the following key objectives, which collectively aim to examine the challenges and prospects of explainability in AFV:

- Provide a comprehensive overview of the current state of explainability in AFV.
- Identify the challenges in the current landscape of AFV explainability, including the concepts of local and global explainability.
- Highlight the importance of creating Explanation-Learning-Friendly (ELF) datasets to advance research in AFV explainability.
- Propose future research directions, including a balanced approach to explainability.
- Contribute to bridging the gap between AFV and XAI principles, and ultimately enhancing the transparency and accountability of AFV systems.

3.1.1 Shortcomings of the Previous Reviews

These few prominent studies Guo et al. (2022); Kotonya and Toni (2020a); Wiegreffe and Marasovic (2021) encompass the relevant research conducted in the areas of explainability in AFV and explainable NLP. Wiegreffe and Marasovic (2021) primarily focuses on explainable NLP datasets and collection methodologies for textual explanations in the context of explainable NLP, rather than AFV. Guo et al. (2022) explores automated fact-checking and discusses the existing datasets and the models, but primarily focus on its applications and challenges rather than explainability. Kotonya and Toni (2020a) provide an in-depth survey of the explanation functionality within fact-checking systems, yet their emphasis remains on generating fact-checking explanations rather than reviewing the broader landscape of AFV. These studies, while valuable in their respective domains, exhibit limitations in offering a comprehensive overview of explainability within AFV systems. As this review progresses, the objective is to address these gaps by conducting a comprehensive analysis of the literature concerning explainability in AFV systems, highlighting

both the strengths and limitations of existing methodologies and datasets. In this process, the paper contributes significantly to a more profound comprehension of vital aspects within the realm of explainable AFV, particularly in the context of XAI.

3.2 Explainable Artificial Intelligence

The field of interpretability in artificial intelligence is experiencing rapid growth, with numerous studies (Ali et al., 2023; Gunning et al., 2021; Kim, 2018) exploring different facets of interpretation. These investigations are often conducted under the umbrella of Explainable AI (XAI), which encompasses various approaches and methodologies aimed at providing explanations for the decision-making process of black-box AI models, providing information on how they generate their outcomes. This section reviews pertinent literature on XAI, with three main objectives: What is explainability? Why is it needed? and How can it be implemented?

The primary objective of XAI is to build models that humans can interpret effectively, especially in sensitive sectors like military, banking, and healthcare. These domains rely on the expertise of specialists to solve problems more efficiently, while also seeking meaningful outputs to understand and trust the solutions provided (Doshi-Velez and Kim, 2017). Additionally, it benefits both domain specialists and developers when appropriate outputs are available, as it encourages investigation into the system when discrepancies occur. However, many AI systems that support decision-making are developed as opaque structures, often referred to as 'black-box' models, which conceal their internal logic from the user and raise practical and ethical concerns (Guidotti et al., 2018; Moradi and Samwald, 2021). These black-box models, while typically offering higher accuracy, do so at the cost of transparency, creating an inherent tension with the need for explainability in machine learning systems (Gunning et al., 2021). In contrast, white-box models are deliberately designed to be interpretable, which makes their outputs easier to understand, but the drawback of these models is the compromise on accuracy. Gray-box models attempt to strike a balance between interpretability and accuracy, offering a favorable trade-off (Ali et al., 2023). Figure 3.1 demonstrates a comparison of these models and briefly conveys the idea of explainability in terms of AI systems. However, it is worth acknowledging that in certain scenarios, such as those involving structured data with naturally meaningful features, simpler classifiers, such as logistic regression or decision lists, may produce competitive results after appropriate preprocessing, as emphasized in the work by Rudin (2019).

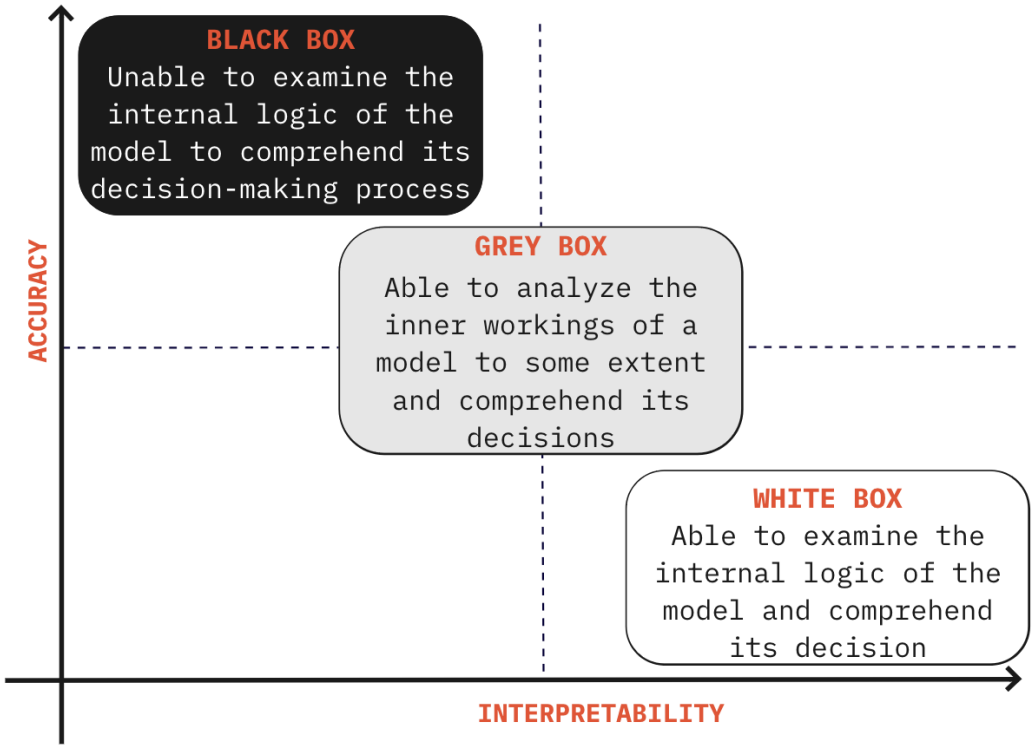


Fig. 3.1 Comparative Analysis of Black-Box, Gray-Box, and White-Box Machine Learning Models, Highlighting an Apparent Trade-off between Accuracy (Y-Axis) and Interpretability (X-Axis)

Upon investigating the reasons for the increase in popularity of this research field, it is evident that XAI has received increasing attention from both academia and industry (Ali et al., 2023; Došilović et al., 2018; Kim, 2018; Kotonya and Toni, 2020a) attributing an inflection point in the middle of the last decade (Gunning et al., 2021). This paragraph presents a brief analysis of the factors contributing to this surge in research interest in an attempt to resolve why Explainability is important and continues to be a pressing requirement in AFV or AI in general. According to studies (Ali et al., 2023; Mueller et al., 2019), as AI becomes more widely implemented, concerns about AI's black-box working paradigm have also become prevalent among governments and the general public. This has led to the need for regulatory authorities to act towards a push for some form of explainability. An initial step towards this AI regulation was taken by the European Parliament in 2016, when it adopted the General Data Protection Regulation (GDPR)[3]. With GDPR

[3]Regulation (EU) 2016/679 of the European Parliament and of the Council, of 27 April 2016., uri=CELEX:32016R0679

policy requiring citizens to receive explanations for algorithmic decisions, explainability has become a significant aspect of algorithm design thereafter (Ali et al., 2023; Goodman and Flaxman, 2017; Gunning et al., 2021). Another authoritative supervision on XAI practices was by the R&D Agency of the United States Department of Defense[4], the Defense Advanced Research Projects Agency (DARPA) (Gunning, 2016). DARPA regulated, an XAI research program, and funded 11 research groups from the USA; they worked in the direction of a common conception; AI systems need to be more explainable to be better understood, trusted, and controlled. The main contribution of this DARPA XAI initiative is the creation of an XAI Toolkit, consolidating the diverse artifacts of the program (such as code, papers, reports, etc.) and the lessons learned from the 4-year program into a centralized publicly accessible repository[5]. This DARPA XAI program and the GDPR policy of the EU Parliament, along with the introduction of the EU AI Act (proposed European law on Artificial Intelligence)[6] contributed majorly to the explainable AI movement observed today (Ali et al., 2023). The drive for greater transparency and accountability in AI is not limited to the global stage; it is also reflected in national initiatives, including those spearheaded by the New Zealand government. For example, various initiatives and intergovernmental standards have been adopted by the New Zealand government, to address the transparency and accountability of AI algorithms. The G20 AI Principles, endorsed by Leaders in June 2019 and based on the OECD AI Policy Observatory, serve as a framework to promote responsible AI use. Similarly, the `Algorithm Charter for Aotearoa New Zealand' aligns with the OECD principles and aims to improve transparency and accountability in AI algorithm usage. The country also actively contributes to global efforts through the OECD portal[7], which showcases AI policy initiatives worldwide and emphasizes the importance of AI transparency. The AI Forum NZ has released its own set of principles, including a focus on AI transparency, while other measures such as the New Zealand Pilot Project from the World Economic Forum[8] further support the objective of improving AI transparency and explainability.

After defining explainability and its significance, the subsequent text in this section explores the various objectives and approaches for XAI implementation, as outlined in the current literature. Following this, it delves into the taxonomy of XAI.

[4]https://www.darpa.mil/program/explainable-artificial-intelligence
[5]https://xaitk.org/
[6]https://artificialintelligenceact.eu/
[7]https://oecd.ai/en/dashboards/overview
[8]https://www3.weforum.org/docs/WEF_Reimagining_Regulation_Age_AI_2020.pdf

3.2.1 XAI Objectives

The current body of work on XAI (Ali et al., 2023; Doshi-Velez and Kim, 2017; Guidotti et al., 2018) identifies several essential conditions necessary for the implementation of explainable AI models. These are categorized into interpretability, accuracy, and fidelity.

Interpretability refers to the degree to which a model and its predictions can be understood by humans. It is often gauged by the complexity of the model, with simpler models generally being more interpretable. Accuracy denotes the model's ability to correctly predict outcomes, particularly for new, unseen instances. It is quantified using metrics such as the accuracy score and the F1-score. Fidelity assesses how well an interpretable model replicates the behavior of a corresponding complex, black-box system. It is crucial for ensuring that the simplified interpretations of the model's decisions are faithful to the original model's logic. Fidelity is evaluated by comparing the interpretable model's predictions to those of the black-box model, often using the same metrics as for accuracy.

The next section reviews the standard approaches suggested in the literature to achieve these XAI objectives.

3.2.2 XAI Approaches

Based on a review of the relevant literature, including studies by various researchers (Ali et al., 2023; Guidotti et al., 2018; Mueller et al., 2019; Murdoch et al., 2019) the practical approaches to achieving the aforementioned XAI objectives are generally categorized into model explainability and post hoc explainability; the latter is further appropriated into interpretations at the prediction level and at the data set level (Murdoch et al., 2019). It is important to note that the model explainability is interchangeably referred to as interpretability in certain works, such as Atanasova et al. (2020).

Model-based explainability

Model-based explainability methods involve the creation of simple and transparent AI models that can be easily understood and interpreted. Such methods are particularly useful when the underlying data relationships are not highly complex, allowing simple models to effectively capture data patterns (Moradi and Samwald, 2021). Models like decision trees and linear regression (unlike Deep Neural Networks), inherently possess model explainability (Gunning et al., 2021). Models

that are inherently interpretable provide their own faithful explanations, accurately representing the computations within the model (Rudin, 2019). However, when dealing with data that exhibit higher degrees of complexity or non-linearity, more intricate black-box models are designed and implemented. In such cases, post hoc explainability techniques are used to extract information about the relationships learned by the model (Ali et al., 2023; Moradi and Samwald, 2021; Shu et al., 2019).

Post hoc explainability

A post hoc explainability method operates on a trained and/or tested AI model, generating approximations of the model's internal workings and decision logic (Moradi and Samwald, 2021; Murdoch et al., 2019). Post hoc methods aim to reveal relationships between feature values and the model's predictions, without requiring access to its internal mechanisms. This enables users to identify the most crucial features in a machine learning task, quantify their importance, reproduce decisions made by the black-box model, and uncover potential biases in the model or the underlying data. The prediction-level interpretation methods revolve around explaining the rationale behind the individual predictions made by the models. These methods delve into identifying the specific features and interactions that contributed to a particular prediction. For example, local interpretability might clarify why a specific loan application was rejected by highlighting the contributing factors (Doshi-Velez and Kim, 2017). In the literature, this concept is referred to by various terms such as local interpretations (Doshi-Velez and Kim, 2017) or local explanations (Atanasova et al., 2020) or local fidelity (Ribeiro et al., 2016). On the other hand, the approaches at the data set level focus on comprehending the broader associations and patterns that the model has learned, with the aim of discerning the patterns related to the predicted responses on a global scale (Murdoch et al., 2019), such as understanding key features governing galaxy formation in astrophysics (Doshi-Velez and Kim, 2017).

3.2.3 XAI Taxonomy

Taxonomy in XAI provides a structured framework to categorize and organize methods, techniques, and approaches for explaining AI and machine learning models, facilitating systematic discussions of model explainability and interpretability. Figure 3.2 provides a hierarchical visualization of XAI by offering a structured view of the XAI approaches mentioned in Section 3.2.2.

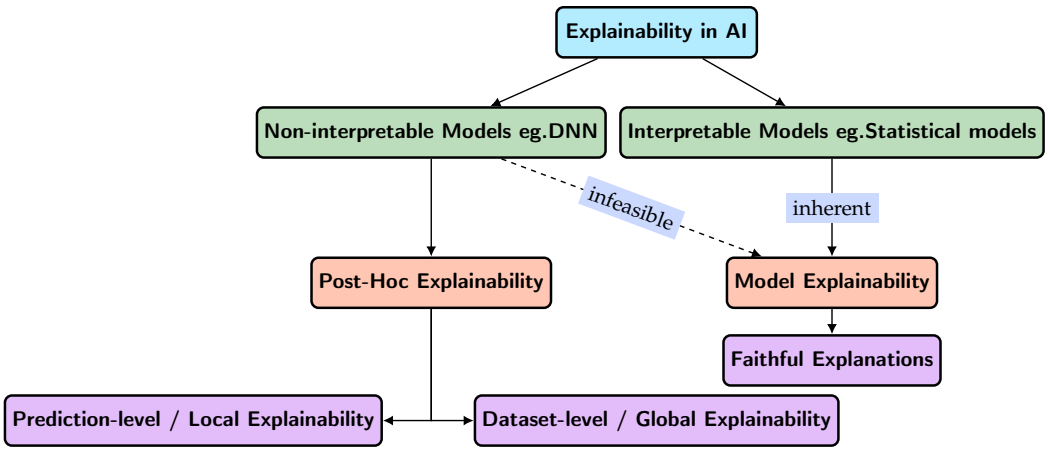


Fig. 3.2 Hierarchical overview of approaches to achieving explainability in AI, focusing on black-box and white-box strategies. The arrows labeled `inherent' and `infeasible' respectively signify the natural transparency of white-box models such as Statistical models, and the difficulties in attaining model explainability in black-box models such as DNNs.

On local explainability, it is worth noting that while an explanation may not achieve complete faithfulness unless it provides a full model description, it is imperative for it to at least achieve local faithfulness. This means that the explanation must accurately correspond to the model's behavior in (at least) the proximity of the instance being predicted, ensuring its meaningfulness. However, it's crucial to highlight that while global fidelity (globally faithful explanations) would encompass local fidelity, local fidelity does not imply global fidelity. Features that have global importance may not necessarily be significant in the local context, and hence the search for globally faithful, but interpretable explanations remains a challenging endeavor, especially when dealing with complex models (Ribeiro et al., 2016).

Section 3.2, presented an overview of explainability concepts, their significance, and associated terminology. The following section will delve into their application within the context of AFV models.

3.3 Explainable AFV

Despite notable progress in the development of explainable AI techniques, achieving comprehensive global explainability in AFV models remains a challenging task. However, this issue encompasses multiple aspects that pose significant obstacles to research in the field of explainable AFV. First, only a relatively small number of automated fact-checking systems include explainability components. Second,

Explainable AFV systems currently do not possess the capability of global explainability. Finally, the existing datasets for AFV suffer from a lack of explanations. This study addresses these factors through three main perspectives: architectural, methodological, and data-based; the examination is conducted in alignment with the objectives and approaches of explainability discussed in the previous section. The emphasis is placed on the data perspective (Section 3.3.3) due to its crucial role in achieving global explainability within the context of AFV. The significance of this perspective is derived from the training dataset's influence on AI model behavior and the critical necessity to achieve a high level of data explainability in AFV.

3.3.1 Architectural Perspective

Majority of AFV systems broadly adopt a three-stage pipeline architecture similar to the **F**act **E**xtraction and **VER**ification (FEVER) shared task (Thorne et al., 2018b), as identified and commented on by many researchers (Chen et al., 2022b; DeHaven and Scott, 2023; Jiang et al., 2021; Krishna et al., 2022; Soleimani et al., 2019; Thorne et al., 2018b; Zhong et al., 2020). These three stages (also called sub-tasks) are, document retrieval (evidence retrieval), sentence selection (evidence selection) and Recognizing Textual Entailment or RTE (label/veracity prediction). The document retrieval component is responsible for gathering relevant documents from a knowledge base, such as Wikipedia, based on a given query. The sentence-retrieval component then selects the most pertinent evidence sentences from the retrieved documents. Lastly, the RTE component predicts the entailment relationship between the query and the retrieved evidence. While the above framework is generally followed in AFV, alternative approaches incorporate additional distinct components to identify credible claims and provide justifications for label predictions, as shown in Figure 3.3. The inclusion of a justification component in such alternative approaches contributes to the system's capacity for explainability within the AFV paradigm.

The majority of AFV systems are highly dependent on deep neural networks (DNNs) for the label prediction task (Kotonya and Toni, 2020a). Furthermore, in recent years, deep learning-based approaches have demonstrated exceptional performance in detecting fake news (Huang et al., 2021). As mentioned in Section 3.2, there is, however, an inherent conflict between the performance of AI models and their ability to explain how they make decisions. However, although existing AFV systems lack inherent explainability (Kotonya and Toni, 2020a), it would be foolish to overlook the potential to use these less interpretable deep models for AFV, as these models possess the ability to achieve state-of-the-art results with a remarkable level

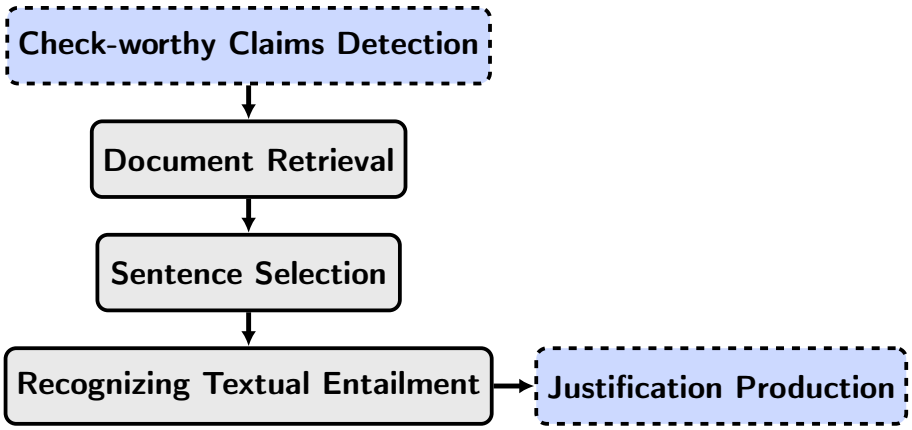


Fig. 3.3 Overview of Stages in Automated Fact Verification: This figure depicts the primary stages — Document Retrieval, Sentence Selection, and Recognition of Textual Entailment, along with optional components for assessing the check-worthiness of claims and providing justifications.

of prediction accuracy. However, this also indicates that model-based interpretation approaches may not be a suitable solution for AFV systems. The reason being that these methods require the involvement of simple and transparent AI models that can be easily understood and interpreted.

Therefore, considering the architectural characteristics of state-of-the-art AFV systems, a potential trade-off solution to achieve explainability may involve incorporating post hoc measures of explainability, either at the prediction-level or dataset-level, while still leveraging the capabilities of less interpretable deep transformer models. The subsequent subsections delve into the attempts made in the literature to incorporate post-hoc explainability in terms of methods and input within the context of AFV.

3.3.2 Methodological Perspective

The methodological aspect looks at the different approaches utilized in existing literature to develop explainable AFV systems.

Summarization Approach

In AFV, extractive and abstractive explanations serve as two types of summarization methodologies, providing a summary along with the predicted label as a form of justification or explanation. Extractive explanations involve directly extracting relevant information or components from the input data that contribute to

the prediction or fact-checking outcome. These explanations typically rely on the emphasis of specific words, phrases, or evidence within the input. On the other hand, abstractive explanations involve generating novel explanations that may not be explicitly present in the input data. These explanations focus on capturing the essence or key points of the prediction or fact-checking decision by generating new text that conveys the rationale or reasoning behind the outcome. It is important to note that terminology can vary across fields. For instance, in the Explainable Natural Language Processing (Explainable NLP) literature, (Wiegreffe and Marasovic, 2021) refers to extractive explanations as `Highlights' and abstractive explanations as `Free-text explanations'.

The approaches to explainability employed by existing explainable AFV systems are primarily extractive. For example, the work of Atanasova et al. (2020) presents the first investigation on the generation of explanations automatically based on the available claim context and utilized the transformer model architecture for extraction summarization purposes. Two models are trained with the intention of addressing this issue. One model focuses on generating post hoc explanations, where the predictive and explanation models are trained independently; while the other model is trained jointly to handle both tasks simultaneously. The model that trains the explainer separately tends to slightly outperform the model trained jointly. Kotonya and Toni (2020b) also approach the task of explanation generation as a form of summarization. However, their methodology differs from that of Atanasova et al. (2020). Specifically, the explanation models of Kotonya and Toni (2020b) are fine-tuned for extractive and abstractive summarization, with the aim of generating novel explanations that go beyond mere extractive summaries. By training the models on a combination of extractive and abstractive summarization tasks, they enabled the models to generate more comprehensive and insightful explanations by leveraging both existing information in the input and generating new text to convey the reasoning behind the fact-checking outcomes.

A potential concern is that these models (both extractive and abstractive) may generate explanations that, while plausible in relation to the decision, do not accurately reflect the actual veracity prediction process. This issue is particularly problematic in the case of abstractive models, as they can generate misleading justifications due to the possibility of hallucinations (Guo et al., 2022).

Logic-based Approach

In logic-based explainability, the focus is on capturing the logical relationships and dependencies between various pieces of information involved in fact verification. This includes representing knowledge in the form of logical axioms, rules, and constraints to provide justifications for the verification results. Chen et al. (2022a) and Krishna et al. (2022) are examples of recent studies that focus on the explainability of fact verification using logic-based approaches.

Chen et al. (2022a) propose a logic-regularized reasoning framework, LOREN, for fact verification. By incorporating logical rules and constraints, LOREN ensures that the reasoning process adheres to logical principles, improving the transparency and interpretability of the fact-verification system. The experimental results demonstrate the effectiveness of LOREN in achieving an explainable fact verification. Similarly, Krishna et al. (2022) highlights the potential of natural logic theorem proving, as a promising approach for explainable fact verification systems. The system named ProoFVer applies logical inference rules to derive conclusions based on given premises, providing transparent explanations for the verification process. The experimental evaluation shows the efficacy of ProoFVer in accurately verifying factual claims while also offering interpretable justifications through the logical reasoning steps.

It is important to acknowledge certain limitations and drawbacks associated with this logic-based approach. First, the complexity and computational cost of logic-based reasoning can limit its scalability and practical applicability in real-world fact verification scenarios. Furthermore, while logic provides a structured and interpretable framework for reasoning, it may not capture all the nuances and complexities of natural language and real-world information. That means, the effectiveness of these approaches heavily relies on the adequacy and comprehensiveness of the predefined logical rules, which may not cover all possible scenarios and domains. Lastly, interpretability of the generated explanations may still be challenging for non-expert users. They may involve complex logical steps that require expertise to fully understand and interpret.

Attention-based Approach

Different from the summarization and the logic-based techniques, explainable AFV systems such as Popat et al. (2018); Shu et al. (2019) use visualizations to illustrate important features or evidence utilized by AFV models for predictions.

This provides users with a means to understand the relationships that influence the decision-making process. For example, the AFV model proposed by Popat et al. (2018) introduces an attention mechanism that directs the focus towards the salient words in the article in relation to the claim. This enables the generation of the most significant words in the article as evidence (words with more weights are highlighted with darker shades in the verdict) and Popat et al. (2018) claim that this strategy enhances the transparency and interpretability of the model. The explanation module of the fact checking framework by Shu et al. (2019) also utilizes the attention mechanism to generate explanations for the model's predictions, highlighting the important features and evidence used for classification.

However, Guo et al. (2022) illustrated several critical concerns associated with the reliability of attention as an explanatory method, citing pertinent studies (Jain and Wallace, 2019; Pruthi et al., 2020; Serrano and Smith, 2019) to reinforce the argument. The authors point out that the removal of tokens assigned high attention scores does not invariably affect the model's predictions, illustrating that some tokens, despite their high scores, may not be pivotal. On the contrary, certain tokens with lower scores have been found to be crucial for accurate model predictions. These observations collectively indicate a possible `fidelity' issue in the explanations yielded by attention mechanisms questioning the reliability and interpretability of attention mechanisms in models. Furthermore, Guo et al. (2022) argue that the complexity of these attention-based explanations can pose substantial challenges for people lacking an in-depth understanding of the model architecture, compromising readability and overall comprehension. This scrutiny of the limitations inherent to attention-based explainability methods highlights the pressing need to reevaluate their applicability and reliability within the realm of AFV.

Counterfactual Approach

Counterfactual explanations, also known as inverse classification, describe minimal changes to input variables that would lead to an opposite prediction, offering the potential for recourse in decision-making processes (Rudin, 2019). These explanations allow users to understand what modifications are needed to reverse a prediction made by a model. In the context of AFV, counterfactual explanations have been explored. The study by Dai et al. (2022), for example, explicitly focuses on the interpretability aspect of counterfactual explanations, in order to help users understand why a specific piece of news was identified as fake. The comprehensive method introduced in that work involves question-answering and entailment

reasoning to generate counterfactual explanations, which could enhance users' understanding of model predictions in AFV. In a recent study (Xu et al., 2023) exploring debiasing for fact verification, the researchers propose a method called CLEVER that operates from a counterfactual perspective to mitigate biases in predicting the veracity. CLEVER stands out by training separate models for claim-evidence fusion and claim-only prediction, allowing the unbiased aspects of predictions to be highlighted. This method could be explored further in the context of explainability in AFV, as it allows users to discern the factors that lead to specific predictions, even if the main emphasis of the cited work was on bias mitigation.

Nevertheless, counterfactual explanations in AFV, while providing valuable insights into why a model makes specific predictions, may also confront challenges in their practical application. One notable limitation lies in the potential complexity and difficulty of interpreting minimal changes in input variables, especially in cases involving complex facts and evidence. This complexity could pose challenges to users in grasping the precise factors that influence the predictions of the models, which is a key aspect in achieving a broader interpretability in AI systems, as discussed in Section 3.2.

Table 3.1 categorizes the existing approaches to develop explainable AFV systems into four methodological aspects discussed: Summarization, logic-based, attention-based, and counterfactual. Each category is illustrated with examples of studies that employ these methods, highlighting their unique contributions as well as potential drawbacks. The table serves as a comprehensive overview, aiding in understanding the various techniques used to enhance the explainability and interpretability in state-of-the-art AFV systems.

Additionally, it is important to note the inherent complexity in typical DNN-based AFV systems. When considered alongside the objectives of XAI outlined in Section 3.2.1, which emphasize that the interpretability of a predictive model is often assessed through its complexity (commonly measured by its size), this factor adds another layer of complexity to the already challenging task of achieving model explainability in state-of-the-art AFVs. However, the situation is equally challenging when it comes to post hoc explainability, especially in terms of achieving global explainability. None of the explainable AFV systems discussed provides global explainability; they mainly focus on prediction-level or local explainability by explaining the model's decision-making process for specific instances or cases. On the other hand, global interpretability at the data set level aims to uncover more general relationships learned by the model and provides a greater understanding

of how the model learns and generalizes across different examples (Murdoch et al., 2019). The following section explores the extent of dataset-level explainability in AFV, leading to an examination of its current state.

Table 3.1 Comparative Analysis of Diverse Methodologies Employed in Enhancing Explainability in Automated Fact Verification Systems

Methodological Aspect	**Examples**	**Drawbacks**
Summarization Approach	Atanasova et al. (2020) Utilizes the transformer model for extractive summarization. Two models trained separately and jointly. Kotonya and Toni (2020b) Fine-tuned for both extractive and abstractive summarization.	May generate misleading or inaccurate explanations. Particularly problematic for abstractive models.
Logic-based Approach	Chen et al. (2022a) LOREN framework uses logical rules for transparency. Krishna et al. (2022) ProoFVer uses natural logic theorem proving.	Complexity and computational cost limit scalability. May not capture all nuances of natural language.
Attention-based Approach	Popat et al. (2018) Uses attention mechanism to focus on salient words. Shu et al. (2019) Utilizes attention for feature and evidence highlighting.	Relies on human experts for visualizations, diverging from the principles of XAI.
Counterfactual Approach	Dai et al. (2022) Focuses on interpretability by generating counterfactual explanations in AFV through question-answering and entailment reasoning. Xu et al. (2023) Proposes the CLEVER method, which operates from a counterfactual perspective to mitigate biases in veracity prediction within AFV.	May face complexity in interpreting minimal input changes, particularly in intricate factual claims and evidence scenarios, potentially hindering broader interpretability.

3.3.3 Data Perspective

The potential of data explainability lies in its ability to provide deep insights that enhance the explainability of AI systems (which rely heavily on data for knowledge acquisition) (Ali et al., 2023; Guidotti et al., 2018). Data explainability methods encompass a collection of techniques aimed at better comprehending the data sets used in the training and design of AI models (Ali et al., 2023). The importance of a training data set in shaping the behavior of AI models highlights the need to achieve a high level of data explainability. Therefore, it is crucial to note that constructing a high-performing and explainable model requires a high-quality training dataset. In AFV, the nature of this dataset, also known as the source of evidence, has evolved over time. Initially, the evidence was primarily based on claims, where information directly related to the claim was used for verification. Subsequently, knowledge-base-based approaches were introduced, utilizing structured knowledge sources to support the verification process. Further advances led to the adoption of text-based evidence, where relevant textual sources were used for verification. In recent developments, there has been a shift towards dynamically retrieved sentences, where the system dynamically retrieves and selects sentences that are most relevant to the claim for verification purposes. The subsequent text examines the implications of these changes through the lens of explainability.

Systems such as Rashkin et al. (2017) that process the claim itself, using no other source of information as evidence, can be termed as `knowledge-free' or `retrieval-free' systems. In these systems, the linguistic characteristics of the claim are considered the deciding factor. For example, claims that contain a misleading phrase are labeled 'Mostly False'. Wang (2017) also employ a similar approach, focusing on linguistic patterns, but incorporate a hybrid methodology by including claim-related metadata with the input text to the deep learning model. These additional data include information such as the claim reporter's profile and the media source where the claim is published. These knowledge-free systems face limitations in their performance, as they depend only on the information inherent in the claim and do not consider the current state of affairs (Thorne and Vlachos, 2018). The absence of contextual understanding and the inability to incorporate external information make dataset-level explainability infeasible in these systems.

In knowledge-base-based fact-verification systems (Bordes et al., 2013; Gardner and Mitchell, 2015; Shi and Weninger, 2016), a claim is verified against the RDF triples present in a knowledge graph. The veracity of the claim is calculated by assessing the error between the claim and the triples based on different approaches

such as rule-based, subgraph-based, or an embedding-based one. The drawback of such systems is the likelihood of a claim being verified as false, based on the assumption that the supporting facts of a true claim are already present in the graph, which is not always feasible. This limited scalability and the inability to capture nuanced information hinder the achievement of explainability in these type of fact verification models.

Unlike the latter two approaches; in the evidence retrieval approach, supporting pieces of evidence for the claim verdict have to be fetched from a relevant source using an information retrieval method. While the benefits of such systems outweigh the limitations of static approaches mentioned earlier, there are certain significant constraints that can also affect the explainability of these models. While the quality of the source (biased or unreliable), availability of the source (geographical or language restrictions), and resources for the retrieval process (time-consuming, and expensive human and computational resources) can have a significant impact on the evidence retrieval and limit the scope of evidence; a deep understanding of claim context is critical to avoid misinterpreted and incomplete evidence which lead to erroneous verdicts. Nevertheless, these limitations suggest that the evidence retrieval approach might not be entirely consistent with key XAI principles such as `Accuracy' and `Fidelity'. This, in turn, casts doubt on the effectiveness of any post hoc explainability measures attempted within this data aspect.

An alternative approach is using text from verified sources of information as evidence; Encyclopedia articles, journals, Wikipedia, and fact-checked databases are some examples. Since Wikipedia is an open source web-based encyclopedia and contains articles on a wide range of topics, it is consistently considered an important source of information for many applications, including economic development (Sheehan et al., 2019), education (Brailas et al., 2015), data mining (Schwenk et al., 2021), and AFV. For example, the FEVER task (Thorne et al., 2018b), an application in AFV, relies on the retrieval of evidence from Wikipedia pages. In the FEVER dataset, each SUPPORTED/REFUTED claim is annotated with evidence from Wikipedia. This evidence could be a single sentence, multiple sentences, or a composition of evidence from multiple sentences, sourced from the same page or multiple pages of Wikipedia. This approach aligns well with the XAI principle of `Interpretability', as Wikipedia is a widely accessible and easily understandable source of information. However, it is crucial to note that Wikipedia also comes with limitations that could impact the `Accuracy' and `Fidelity' principles of XAI, which can potentially impact the interpretability of models relying on Wikipedia as a primary data source. Firstly,

like any other source, Wikipedia pages can contain biased and inaccurate content, and these can remain undetected for a longer period (same with outdated information); this compromises the `Accuracy' of any AFV model trained on these data. Secondly, despite covering a wide range of topics, Wikipedia suffers deficiencies in comprehensiveness[9], limiting a model's ability to understand contextual information fully, thereby affecting `Interpretability. Lastly, models trained predominantly on Wikipedia's textual content can develop biases and limitations inherent to the nature and scope of Wikipedia's content, impacting both `Fidelity' and `Interpretability' when applied to diverse real-world scenarios and varied types of unstructured data.

Given these considerations and their misalignment with the XAI objectives of `Interpretability,' `Accuracy', and `Fidelity,' it becomes evident that relying solely on Wikipedia as a training dataset may not be the most effective pathway toward explainable AFV.

Alternatively, Wikipedia can be used as an elementary corpus to train the AI model to achieve a general understanding of various knowledge domains for AFV, and this background or prior knowledge can then be harnessed further with additional domain data to gain a deeper context (which helps the model to attain information on global relationship and thus increase explainability). Being the largest Wikipedia-based benchmark dataset for fact verification (Shorten et al., 2021; Zhong et al., 2020), the FEVER dataset can unarguably be considered as this elementary corpus for AFV tasks, and Transformers and Transfer Learning is the most pragmatic technology choice for AFV according to state-of-the-art systems (DeHaven and Scott, 2023; Krishna et al., 2022; Stammbach, 2021).

The quality of the data set used or created for an application is a major factor in determining the explainability of a transformer-based AFV model and its ability to comprehend the underlying context. For example; Wadden et al. (2020) developed the SCIFACT data set in order to expand the ideas of FEVER to COVID-19 applications. SCIFACT comprises 1.4K expert-written scientific claims along with 5K+ abstracts (from different scientific articles) that either support or refute each claim and is annotated with rationales, which consists of a minimal collection of sentences from the abstract that imply the claim. The study demonstrated the obvious advantages of using such a domain-specific dataset (can also be called subdomain here since scientific claim verification is a sub task of claim verification) as opposed to just using a Wikipedia-based evidence dataset. Wadden et al. (2020) argues that the inclusion of rationales in the training data set "facilitates the development of

[9]https://en.wikipedia.org/wiki/Reliability_of_Wikipedia#Coverage

interpretable models" that not only label predictions but also identify the specific sentences necessary to support their decisions. However, the limited scale of the dataset, consisting of only 1.4K claims, necessitates caution in interpreting assessments of system performance and underscores the need for more expansive datasets to propel advancements in explainable fact-checking research.

Building on this perspective of improving the quality and diversity of the data set, Hanselowski et al. (2019) critically evaluated the FEVER corpus, emphasizing its reliance on synthetic claims from Wikipedia and advocating for a corpus that incorporates natural claims from a variety of web sources. In response to this identified need, they introduced a new, mixed-domain corpus, which includes domains like blogs, news, and social media, the mediums often responsible for the spread of unreliable information. This corpus, which encompasses 6,422 validated claims and over 14,000 documents annotated with evidence, addresses the prevalent limitations in existing corpora, including restricted sizes, lack of detailed annotations, and domain confinement. However, through meticulous error analysis, Hanselowski et al. (2019) discovered inherent challenges and biases in claim classification, attributed to the heterogeneous nature of the data and the incorporation of Fine-Grained Evidence (FGE) from unreliable sources. These findings illustrate substantial barriers to realizing the fundamental goals of XAI, particularly accuracy and fidelity. Moreover, Hanselowski et al. (2019)'s focus on diligently modeling meta-information related to evidence and claims could be understood as their implicit recognition of the crucial role of explainability in the realm of automated fact-checking. By suggesting the integration of diverse forms of contextual information and reliability assessments of sources, they highlight the necessity of developing models that are not only more accurate but also capable of providing reasoned and understandable decisions, a pivotal step towards fostering explainability in automated fact-checking systems.

Table 3.2 Comparative Analysis of Dataset Types and Their Impact on Explainability in AFV Systems

Fact_Verificati Dataset Type	**Example_Stud**	**Knowledge_Type**		**Text_Type**		**Domain_Type**		**Remarks**
		Knowledg Free	*External-knowledge*	*Structure Data*	*Free-Text*	*Single-Domain*	*Multi-Domain*	
Knowledge-free Systems	Rashkin et al. (2017), Wang (2017)	✓	×	–	–	–	–	Lack of contextual understanding, dataset-level explainability infeasible
Knowledge-Base-Based	Shi and Weninger (2016), Gardner and Mitchell (2015)	×	×	✓	×	✓	✓	Limited scalability, inability to capture nuanced information
Wikipedia-Based	Thorne et al. (2018a)	×	✓	×	✓	×	✓	Biased and inaccurate content, limited comprehensiveness
Domain(Single Specific-Corpus	Wadden et al. (2020)	×	✓	×	✓	✓	×	Limited size, potential for biased evaluation
Mixed-domain-Corpus(non-Wikipedia-based)	Hanselowski et al. (2019)	×	✓	×	✓	×	✓	Challenges in classification due to heterogeneous data (impact accuracy); evidence from unreliable sources (impact fidelity)

Table 3.2 offers a comprehensive categorization of the datasets used in Fact Verification systems, highlighting a variety of dataset types, each highlighting distinctive attributes and challenges. The datasets are categorized meticulously based on their inherent nature and source, such as `Knowledge-free Systems', `Knowledge-Base-Based', `Wikipedia-Based', `Domain(Single)-Specific-Corpus', and `Mixed-domain-Corpus (non-Wikipedia-based)'. Each type is represented with illustrative studies and remarks to provide insight into the inherent limitations or challenges in relation to enhancing explainability in AFV systems. The categorization is enriched with sub classifications under `Knowledge Type', `Text Type', and `Domain Type'. `Knowledge-free systems' are denoted with dashes (-) under `Text Type' and `Domain Type', indicating the inherent absence of these attributes. This underscores the retrieval-free nature of such systems, which predominantly rely on the intrinsic linguistic features of the claims, thus lacking contextual understanding. The `Knowledge-Base-Based' type can be either single-domain, or multi-domain represented by check marks in both sub-categories under `Domain Type'. This illustrates the versatility of knowledge-based systems in utilizing structured information from a specialized domain or amalgamating insights from multiple domains. The ability to cater to varied domains accentuates the expansive applicability of such systems, though it also brings forth challenges related to scalability and capturing nuanced information. `Wikipedia-Based' datasets, inherently multi-domain, are highlighted separately to focus on the specific challenges of using Wikipedia as the main information source, such as dealing with potential biases and inaccuracies. The `Domain(Single)-Specific-Corpus is distinguished as it focuses on a specialized or singular domain, providing depth and specificity. While this focus allows for a detailed exploration of a particular domain, it also poses limitations due to the restricted scope and potential biases inherent to the selected domain, thereby affecting the overall evaluation and applicability of the system. Additionally, the `Mixed-domain Corpus' type emphasizes the inclusion of diverse domains, especially those not solely reliant on Wikipedia, addressing the challenges arising from data heterogeneity and reliability.

The categorization in Table 3.2, coupled with associated remarks, is intended to act as a resource, providing information on the various challenges and possibilities to improve explainability within AFV systems. This categorization can guide researchers and practitioners in making informed decisions regarding dataset selection and utilization, providing a clearer understanding of the implications and limitations of different dataset types in the context of Automated Fact Verification.

This study acknowledges the extensive investigations conducted by Wiegreffe and Marasovic (2021) in Explainable NLP and by Kotonya and Toni (2020a) in Explainable AFV, which provide meticulous lists and insightful analyses of prevalent datasets in their respective fields. It is crucial to clarify that the endeavor in this section (Section 3.3.3) does not aim to perform an exhaustive review of datasets, a task diligently undertaken by the aforementioned studies. Instead, it is uniquely positioned to illuminate the distinctive attributes and inherent diversity within various dataset types in AFV. This attempt to examine the impact of different data types on explainability serves as a thoughtful addition to ongoing discussions and reflections on the subject, offering a new perspective on the multifaceted interactions between data diversity and explainability in AFV.

3.4 Discussion

While fact-checking datasets commonly support the standard 3-stage pipeline of fact verification, there is currently a lack of datasets that specifically facilitate explanation learning aligned with government and intergovernmental standards on XAI. This is of paramount importance towards Explainable AFV, that if an AI system is expected to produce explanations, it should have the ability or opportunity to consume explanations. To achieve this, it is necessary to train the model network using an explanation-learn-friendly (ELF) dataset. However, prominent large-scale datasets like FEVER (Thorne et al., 2018a) and MultiFC (Augenstein et al., 2019) lack this aspect of the fact verification task. Furthermore, currently there are no alternative resources available to address this limitation, as commented on by Stammbach and Ash (2020). To create an ELF dataset, it is essential to analyze the data set practices of fact verification systems with a focus on explainability. This paper has undertaken this crucial initial step and found that the absence of an explanation-based fact verification corpus presents a significant obstacle to advancing research in the field of explainable fact-checking.

In addition to the lack of suitable ELF datasets in AFV, another significant challenge to the growth of the explainable AFV field is the ambiguity and discrepancies surrounding the concepts of local and global explainability. Global interpretability refers to the ability to comprehend the overall logic and reasoning of a model, including all possible outcomes. It involves understanding the complete decision-making process and the underlying principles of the model. On the other hand, local interpretability refers to the ability to understand the specific reasons or factors that

contribute to a particular decision or prediction. It focuses on the interpretability of individual predictions or decisions rather than the entire model. These terms are not consistently understood and implemented in different research communities, leading to confusion and causing impedimenting progress in the field. While XAI researchers (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Moradi and Samwald, 2021; Murdoch et al., 2019) generally adhere to a consistent understanding of local and global explainability, Explainable AFV researchers have different interpretations and perspectives, contributing to the ambiguity surrounding explainable AFV. For example, Kotonya and Toni (2020b) focuses on local coherence and global coherence, evaluating sentence cohesion and the appropriateness of explanations in relation to the claim and associated label, both at the prediction level. On the other hand, Atanasova et al. (2020) discuss the explainability as providing local explanations for individual data points, without specifically addressing local or global aspects. As a result, the definitions of local coherence, global coherence, and explainability in AFV studies predominantly refer to prediction-level explainability, leaving the concept of global explainability in AFV insufficiently defined.

The lack of recognition of the importance of global explainability is evident in the implementations as well. Existing systems primarily focus on local explainability, which hampers adequate understanding of the system's decision-making process at the data set level. In an extensive survey conducted by Kotonya and Toni (2020a) on explainable AFV systems, it was found that all the examined systems focused primarily on providing explanations for individual predictions rather than offering explanations about the underlying fact-checking model itself. This indicates a prevalent trend in the field of explainable AFV, where the emphasis is on local explainability. However, this local exaplinability is not sufficient for AFV systems because it only provides insights into individual predictions without offering a holistic view of the system's overall behavior. Furthermore, global explainability is crucial for AFV systems, as it provides a comprehensive understanding of how the system arrives at its predictions and decisions. This global approach also allows AFV systems to align with advances in XAI research and comply with the XAI principles, enabling transparency and accountability.

In addition to the ambiguity surrounding local and global concepts, the field of explainable AFV is further complicated due to variations in how explainability concepts are categorized, suggesting a lack of consensus on taxonomy. For example, while Shu et al. (2019) distinguishes between intrinsic explainability and post hoc explainability, other researchers in explainable AFV, such as Atanasova et al.

(2020), propose a categorization that broadly divides XAI into interpretability and explainability. Shu et al. (2019) describe intrinsic explainability as the process of creating self-explanatory models that inherently incorporate explainability. This suggests that their definition of `intrinsic explainability' closely aligns with the general notion of `interpretability' related to model-level reasoning, as discussed in Section 3.2.3. However, the choice of the term `intrinsic' by Shu et al. (2019) adds a distinct nuance to this categorization. On the other hand, their view on post hoc explainability is in line with standard XAI. In contrast, while Atanasova et al. (2020) aligns `interpretability' with the mainstream XAI taxonomy, they adopt a narrower view for `explainability', reserving it for local explanations of individual instances, which is a subset of post-hoc explainability. This deviates from the wider view where `explainability' usually refers to model-level justifications.

These disparities in taxonomy demonstrate that the ambiguity extends beyond the local and global dimensions, contributing to the overall ambiguity within the field of explainable AFV. This disagreement and discrepancy among the relatively few existing explainable AFV systems pose significant challenges for the growth and advancement of research in this field and highlights the need for a more standardized approach to explainability in AFV systems.

3.4.1 Limitations

This study concentrates on exploring the explainability of DNN-based AFV models, consequently not addressing other explainability approaches such as rule discovery (Ahmadi et al., 2019; Gad-Elrab et al., 2019). This research gap provides an opportunity for future studies to investigate the model explainability of DNNs, particularly transformer models, in the context of AFV.

Similarly, to limit the scope of this paper, this study did not address the absence of a clear and established link between the various interpretation methods proposed in the literature and the evaluation criteria for measuring explainability; the lack of clarity regarding how to measure explainability is a significant challenge in this field of the research. This aspect warrants further investigation in future research to enhance the assessment of explainable AFV systems.

3.5 Future Research Directions

In addition to the future plans outlined in the limitations of this study, the following directions for exploration and research are proposed.

- *Direction 1:* Exploring a Balanced Approach to Explainability in AFV: Researchers should explore the development of techniques and methodologies aimed at achieving a balanced approach to explainability, integrating both global and local perspectives in AFV systems. This involves understanding the broader relationships and patterns that underlie AFV model decisions across diverse factual claims and evidence (global explainability), while also addressing the specific concerns related to individual instances (local explainability). For instance, a nuanced exploration of gray-explainability could involve refining gray-box models to optimize the trade-off between interpretability and accuracy, ensuring that the explanations provided are as meaningful and understandable as possible without incurring a substantial loss in predictive accuracy. Although examples such as dispute resolution and individual patient treatment decisions illustrate the broader applicability and importance of this approach beyond the realm of fact verification, they underscore the universal need for tailored and comprehensible explanations in individual cases. In fact verification systems, a balanced approach is particularly crucial for gaining both a localized understanding of individual claims and a broader insight that can inform strategies in handling diverse types of information and evidence. By investigating methods that provide insights into AFV model behavior and reasoning patterns on both the macro and micro levels, researchers can work towards achieving a holistic understanding of explainability in AFV systems. Following the principles of XAI, a potential starting point could be to explain multiple representative individual predictions (locally) as a means to gain insights toward a more comprehensive understanding, as suggested by Ribeiro et al. (2016). This nuanced exploration, which aligns with the overarching goal of achieving explainability in AFV systems, ensures that the insights gained are as widely applicable as they are individually relevant, potentially leading to more informed and equitable decision-making processes across different domains.

- *Direction 2:* Comprehensive Investigation and Comparative Analysis of AFV Datasets: Future research endeavors could benefit from undertaking a meticu-

lous and comprehensive review of the applicable data sets for AFV, informed by the insights provided in the tables 3.1 and 3.2. Table 3.1 outlines a comparative analysis of various methodologies used to improve explainability in AFV systems, while Table 3.2 delves into the distinctive attributes and inherent diversity within various types of data set in AFV. A focused study in this direction could reveal deeper insights into the suitability and compatibility of various datasets with different AFV models and explainability techniques, providing a more nuanced understanding of the interactions between dataset characteristics and explainability. Such an investigation would not only enrich the understanding of the influence of diverse datasets on the explainability of AFV models but also reveal untapped potential in utilizing underexplored types of dataset to enhance model transparency and interpretability. By synergizing the diverse techniques for explainability and the variety of dataset types highlighted in the tables, this research direction has substantial potential to lessen the gap in the field of explainable AFV.

- *Direction 3:* Development of an Explainability-Focused, Explanation Learning-Friendly (ELF) Dataset: As a logical progression from Direction 2, researchers should prioritize developing an ELF dataset to address the lack of explanations in existing AFV datasets, enabling more nuanced studies in explainability in AFV. This customized data set would serve as a benchmark to assess the effectiveness of various AFV models in generating meaningful explanations, thereby fostering advancements in creating explainable AFV systems. Such a focused endeavor would be pivotal in bridging existing gaps and furthering research in explainable AFV, allowing for an exploration of the interplay between dataset attributes, model structures, and explainability methodologies.

3.6 Conclusions

This study addresses the challenge of explainability within AFV systems, shedding light on key insights that both researchers and practitioners in the field can leverage.

Drawing from a comprehensive analysis of AFV models and the principles of XAI, this paper outlines a clear road map for future research. This study firmly advocates for a shift in focus from local explainability which currently dominates AFV systems, towards a broader embrace of global explainability in line with XAI objectives. To

catalyze this transition, the necessity for developing specialized training datasets tailored explicitly for the pursuit of global explainability is highlighted.

The examination of data practices within current AFV frameworks revealed critical gaps and limitations. Moreover, it exposed inconsistencies and discrepancies among AFV systems regarding the concepts and perspectives of explainability.

While this study serves as a foundation for future research, it is imperative to recognize that the path from manual to automated fact verification remains incomplete. The incorporation of explainability as an essential functionality in modern AI systems must be prioritized, as highlighted in the problem statement in the Introduction.

In conclusion, this study makes a valuable contribution to the expanding field of AFV and XAI by offering a determinate, interconnected approach to address the pressing challenge of explainability in AFV systems. It is anticipated that the analysis and insights gained from this paper will catalyze both researchers and practitioners in the field for further research in explainability, ultimately leading to more transparent and accountable AFV systems.

Chapter 4

Prelude - Manuscript 2

One of the key limitations in explainability identified in Manuscript 1 (Chapter 3) is the lack of methodologies that extend beyond isolated claim-level justifications to incorporate broader thematic structures. Addressing this gap is particularly important, as a critical shortcoming in AFV systems is the absence of systematic methods for discovering and representing thematic relationships.

Manuscript 2 (Chapter 5) addresses this gap by introducing a novel approach for thematic discovery and visualization. This study presents the SOI methodology, which identifies semantically related claims and evidence to form structured thematic clusters. By constructing graph-based representations of SOIs, this framework enables a visual exploration of a claim's global context, illustrating its thematic alignment within the broader verification landscape. While this study does not yet integrate SOI into evidence retrieval or claim verification, it provides a foundational step toward context-aware fact verification, later operationalized in Manuscript 3 (Chapter 7). Through this contribution, Manuscript 2 establishes the groundwork for leveraging thematic structures in AFV, bridging the gap between local claim verification and global thematic reasoning.

Chapter 5

Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems (Manuscript 2)

5.1 Introduction

The "AI/ML Black Box Paradox" highlights the challenges of understanding the decision-making processes of advanced AI and machine learning systems, where the complexity of these methods makes their predictions challenging for humans to comprehend. This lack of clarity raises concerns about the reliability and trustworthiness of these systems, especially in sensitive applications. As a result, there has been a growing research interest in the field of explainable AI (XAI) to make AI models more transparent and understandable (Ali et al., 2023; Gunning et al., 2021; Vallayil et al., 2023). XAI encompasses a range of strategies and techniques designed to clarify the decision rationale of AI models by explaining how they generate their results. These efforts may leverage the intrinsic explainability of simpler, interpretable models like statistical models or employ post-hoc explainability methods for more complex models that lack intrinsic explainability, such as deep neural networks (DNNs). In some cases, XAI combines these approaches by using interpretable machine learning models for the post-hoc analysis of black-box models, as exemplified

by LIME (Ribeiro et al., 2016). However, despite being less interpretable, DNNs such as transformers, which excel at managing long dependencies and understanding contextual relationships in sequential data, are widely adopted due to their high accuracy across various real-life applications, including Automated Fact Verification (AFV) systems.

AFV systems are designed to assess the veracity of claims by leveraging advanced AI techniques. Typically, these systems follow a three-stage pipeline: document retrieval (gathering relevant documents), evidence selection (identifying pertinent evidence sentences), and Recognizing Textual Entailment (RTE, which predicts the relationship between the claim and the evidence). Given the reliance on complex models like transformers, as discussed earlier, these AI-based systems often face a trade-off between accuracy and transparency (Ali et al., 2023). However, both these aspects are critical, especially in the case of AFV. To address this, post-hoc explainability techniques are recommended in the literature and are generally integrated into the RTE phase of AFV systems (Atanasova et al., 2020; Kotonya and Toni, 2020b).

Post-hoc methods work by analysing trained AI models to approximate their internal workings and decision logic. These methods reveal relationships between feature values and model predictions without requiring access to the model's internal mechanisms. This enables users to identify crucial features, quantify their importance, reproduce the decisions made by the black-box model, and detect potential biases in the data or model itself. Post-hoc explainability can be categorized into local and global perspectives, which have been discussed with slight variations in the literature (Atanasova et al., 2020; Shu et al., 2019). A widely accepted interpretation is that local explainability pertains to the analysis of individual data points, whereas global explainability focuses on understanding the broader patterns and associations in the model's decisions across the entire dataset (Doshi-Velez and Kim, 2017; Murdoch et al., 2019; Ribeiro et al., 2016).

However, despite the recognized importance of global explainability in AFV, its implementation remains limited, as existing systems primarily focus on local explainability, which only provides insights into individual predictions without offering a holistic view of the system's overall behaviour (Kotonya and Toni, 2020a). This narrow focus on local explainability fails to capture the broader decision-making patterns of AFV models, indicating the need for a more comprehensive approach. To address this gap, we conducted and published a comprehensive literature review (Vallayil et al., 2023). This current study builds on the findings

and gaps identified in that review, focusing on a research direction that advocates integrating both global and local perspectives to provide a balanced approach to explainability in AFV systems. We employ methods that offer insights into AFV model behaviour and reasoning patterns on both global and local levels, aiming for a holistic understanding of explainability in AFV systems. Through this approach, we seek to understand the broader relationships and patterns that inform AFV model decisions while addressing specific concerns related to individual instances.

5.2 Methodology

This section outlines our structured approach to enhancing explainability in AFV systems, focusing on the use of clustering techniques to achieve global explainability. Our methodology encompasses generation of embedded vectors, application of thematic clustering, and the identification of evidence clusters. These steps form the foundation for the subsequent case study and visualizations discussed in Section 5.3.

5.2.1 Dataset Overview

The dataset we employed in this study is novel and specifically curated for XAI research in AFV. Each claim in the dataset is associated with a unique claim ID, indicating its thematic context, such as climate change, COVID-19, and electric vehicles, facilitating structured analysis of these themes and their impact on fact verification. Claims are labelled as True, False, or Not Enough Info, similar to the structure used in standard fact verification datasets like FEVER (Thorne et al., 2018b). Additionally, each claim is paired with six annotated pieces of evidence. By structuring the dataset with multiple pieces of evidence for each claim, we support detailed exploration for XAI research. At the same time, the decision to store these as separate entries, along with the three-class labelling format, aligns with conventional fact verification practices. Thereby, we anticipate that this dataset will support research initiatives in both traditional fact verification as well as explainability-focused studies in AFV systems. Moreover, the dataset's structure is also tailored to facilitate both local and global explainability, which is a key focus of our study. On a local level, the multiple pieces of evidence for each claim allow for a granular examination of the relationships between a claim and its supporting evidences, providing detailed insights into the rationale behind individual claim evaluations.

On a global level, the thematic organization and clustering techniques applied to the dataset help uncover broader patterns and relationships, enhancing the overall understanding of the AFV system's decision-making process.

A detailed description of the dataset, including its creation, annotation processes, and potential applications in AFV systems, will be provided in an upcoming journal publication, with further empirical analysis to be discussed in subsequent work as outlined in our future research section.

5.2.2 Data Preprocessing and Embedded Vector Generation

We begin by loading and preprocessing the dataset, denoted as D, and then extract claims and their associated evidences within a specific theme, such as climate change. This thematic subset, represented as DT , serves as the basis for subsequent analyses. The preprocessing steps ensure that the data is clean, consistent, and ready for embedding generation and clustering. This process involves grouping the data by claims and their corresponding evidences, as well as standardizing text to maintain uniformity. In our research, selecting an effective model for generating sentence vectors is crucial, as accurately capturing the semantic nuances and contextual relationships within the text is vital for deducing the interrelations between claims and other elements in the dataset. Traditional models like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which provide static embeddings, often fall short in capturing the depth needed for nuanced text understanding in NLP tasks like AFV. To address this, we considered BERT (Devlin et al., 2019) and its variants, which generate context-aware vectors and offer a significant improvement over traditional model. However, despite BERT's success in many NLP tasks, its sentence vectors without fine-tuning have been found to be inadequate for tasks requiring detailed semantic textual similarity, often even underperforming compared to static embeddings like GloVe (Li et al., 2020a; Reimers and Gurevych, 2019).

Therefore, we selected Sentence-BERT (SBERT), which modifies the pre-trained BERT network by incorporating siamese and triplet network structures (Reimers and Gurevych, 2019). This approach enables SBERT to produce semantically meaningful sentence vectors that can be effectively compared using cosine similarity. Among the pre-trained SBERT models available, we chose the open-source *all- mpnet-base-v2* for our research, which is fine-tuned on over 1 billion textual pairs for textual similarity tasks. This model has demonstrated superior performance in capturing both semantic and syntactic correspondences among the fine-tuned models for general-purpose textual similarity tasks, specifically those based on pre-trained

transformer models like BERT and MPNet, as well as parameter-reduced models like DistilBERT (Jayanthi et al., 2021), making it particularly effective for AFV tasks that require a nuanced understanding of textual similarity.

The high-quality vectors produced by *all- mpnet-base-v2* are not only critical for accurately assessing textual similarity between claims and evidence in AFV but also support our approach to clustering and visualization. These vectors serve as the foundation for exploring the interconnections within thematic clusters, which is central to enhancing the explainability of AFV systems in our approach, as discussed in the upcoming sections.

5.2.3 Thematic Clustering with GMM-EM

Utilizing the embeddings generated in the previous step, we apply Gaussian Mixture Models (GMM) with the Expectation-Maximization (EM) algorithm to identify thematic clusters in the dataset. Although GMM and EM have been widely used in fields such as speaker identification, emotion recognition, and brain image segmentation (Al-Dujaili Al-Khazraji and Ebrahimi-Moghadam, 2024; Barai et al., 2022; Binti Kasim et al., 2021; Jiao et al., 2023; Li et al., 2020b; Moondra and Chahal, 2023), our approach uniquely adapts these techniques for AFV, modelling the data as a combination of multiple underlying patterns. Each pattern is represented by a Gaussian distribution defined by its mean and spread, helping us to identify distinct thematic clusters within the dataset.

While GMM supports soft clustering, where data points can belong to multiple clusters with varying degrees of membership, our methodology focuses solely on the primary cluster assignment for each data point, bypassing the probabilistic flexibility that GMM offers. This approach ensures that each claim and evidence is clearly associated with a single cluster, simplifying the subsequent steps of inference and visualization. This is particularly advantageous in the context of AFV as it allows for a clear understanding and representation of the relationships between claims and evidences, facilitating more precise analysis and interpretation. The clusters are identified as follows:

$$\mathrm{EM}(T) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(e_{T,i} \mid \theta) \rightarrow \{C_i\}_{i=1}^{k} \tag{5.1}$$

where:

- EM(T) denotes the EM algorithm applied to the embeddings of data within the theme T.
- $e_{T,i}$ is the embedding of the i-th text input (claim or evidence) from the thematic subset $\mathcal{D}_T$.
- θ represents the set of parameters of the GMM, including the means, covariances, and mixture weights of the Gaussian distributions.
- $p(e_{T,i} \mid \theta)$ is the probability density function of the embedding $e_{T,i}$ given the model parameters θ, as defined by the GMM.
- $\{C_i\}_{i=1}^{k}$ represents the set of clusters formed as a result of the EM algorithm, where each C_i is a cluster and k is the total number of clusters.

5.2.4 In-Depth Analysis of Claims Within Thematic Clusters

To conduct a focused analysis on a specific claim c_i within a theme, we programmatically determine the cluster C_i that contains the claim, among the k clusters identified by the EM algorithm (Equation 5.1). Within this cluster, we further refine the embeddings to identify a subset of pertinent evidences and claims based on their semantic relevance to the claim c_i using a similarity measure; this subset hereafter referred to as the Subset of Interest (SOI). The SOI for c_i is defined as follows (in Equation 5.2):

$$\mathcal{S}(C_i) = \{e_{i,j} \mid \text{sim}(e_{i,j}, c_i) > \delta \text{ and } e_{i,j} \in C_i\} \cup \{c_k \mid \text{sim}(c_k, c_i) > \delta \text{ and } c_k \in C_i\} \quad (5.2)$$

where:

- $\mathcal{S}(C_i)$ represents the set of relevant evidences and related claims associated with the claim c_i.
- $e_{i,j}$ denotes the j-th evidence associated with the claim c_i.
- $\text{sim}(e_{i,j}, c_i)$ is the similarity measure between the evidence $e_{i,j}$ and the claim c_i.
- $\text{sim}(c_k, c_i)$ denotes the similarity measure between a claim c_k and the claim c_i.
- δ is the similarity threshold above which an evidence $e_{i,j}$ or a claim c_k is considered relevant to the claim c_i.

This analysis provides the foundational data for the graph-based visualizations described in Section 5.3.2, where we visualize the interconnections and contextual relationships that are specifically influential in the context of the selected claim.

5.3 Case Study Preliminary Findings: Graph-Based Visualizations and Interpretive Insights

In this section, we present preliminary findings from a case study conducted using our novel dataset, demonstrating the implementation of the proposed methodology described in Section 5.2. These interim results demonstrate the effectiveness of thematic clustering in providing meaningful insights through visualization, demonstrating the relationships and interconnections among claims and their associated evidences (direct and indirect) within the dataset. These visualizations provide a macro-level view of the thematic clusters (Figure 5.1) and a micro-level analysis of specific claims and their related evidences (Figure 5.2).

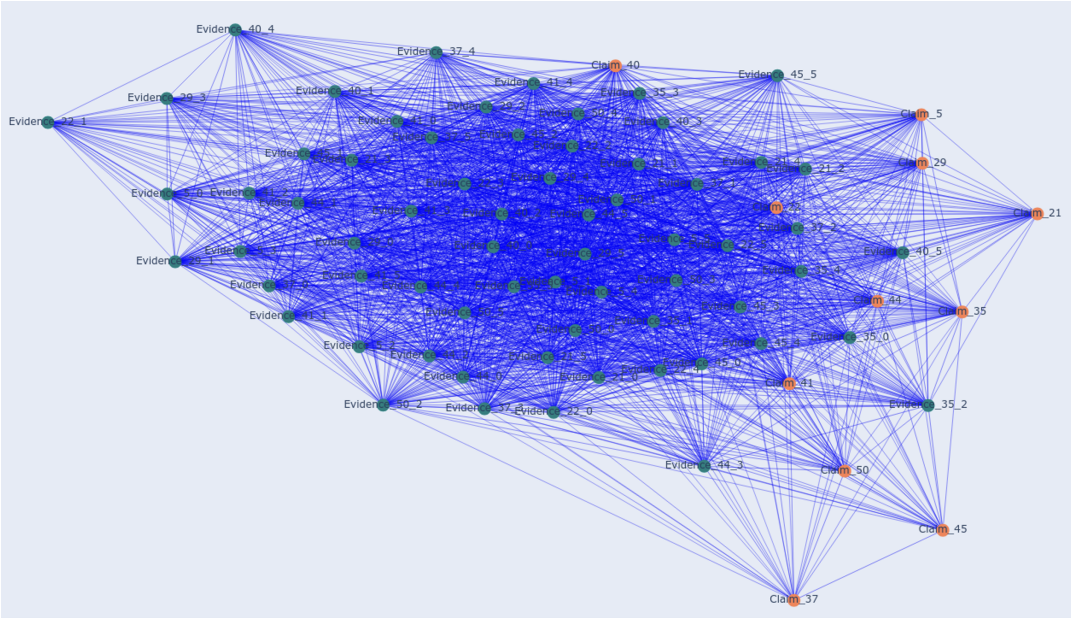


Fig. 5.1 Graph visualization of Cluster 0. Represents the relationships between claims (coral nodes) and evidences (teal nodes) within a specific cluster of a theme. The edges represent the connections between claims and evidences, demonstrating the complexity and density of the thematic cluster. This visualization provides a comprehensive overview of the thematic context, setting the stage for more focused analyses.

5.3.1 Thematic Cluster Visualization

We start by visualizing the thematic clusters identified through the GMM-EM (Section 5.2.3). For this case study, we focus on the 'Climate Change' theme, which has been clustered into three distinct groups (k=3 in Equation 5.1). This resulted in the total number of samples in this theme (n=378) to be segregated to, Cluster 0 with 140 samples, Cluster 1 with 138 samples, and Cluster 2 with 100 samples.

We generated graph visualizations for these clusters, using distinct colours to differentiate between claims and evidence pieces, making the relationships visually clear (Figure 5.1 for the visualization of Cluster 0). The edges between nodes represent the semantic relation between claims and evidence sentences, indicating their relevance. The graph is complex and challenging to interpret in detail due to the dense interconnections. However, it serves as an essential overview of the thematic cluster, providing context for more focused analyses in the following subsection. For example, the claim marked as 'Claim_21,' visible in Cluster 0's visualization at the top-right corner, is explored in detail for in-depth analysis in the next subsection as part of the case study.

5.3.2 In-Depth Visualization of Claims Within Thematic Clusters

To gain insights into the relationships and dependencies of a selected claim from a post-hoc explainability perspective (Moradi and Samwald, 2021), we apply the focused analysis conceptualized in Section 5.2.4, visualizing the SOI identified through Equation 5.2 for Claim_21. This graph highlights the thematic interconnections the claim has with other participants in its cluster, offering a more refined view of its relationships within the SOI (Figure 5.2).

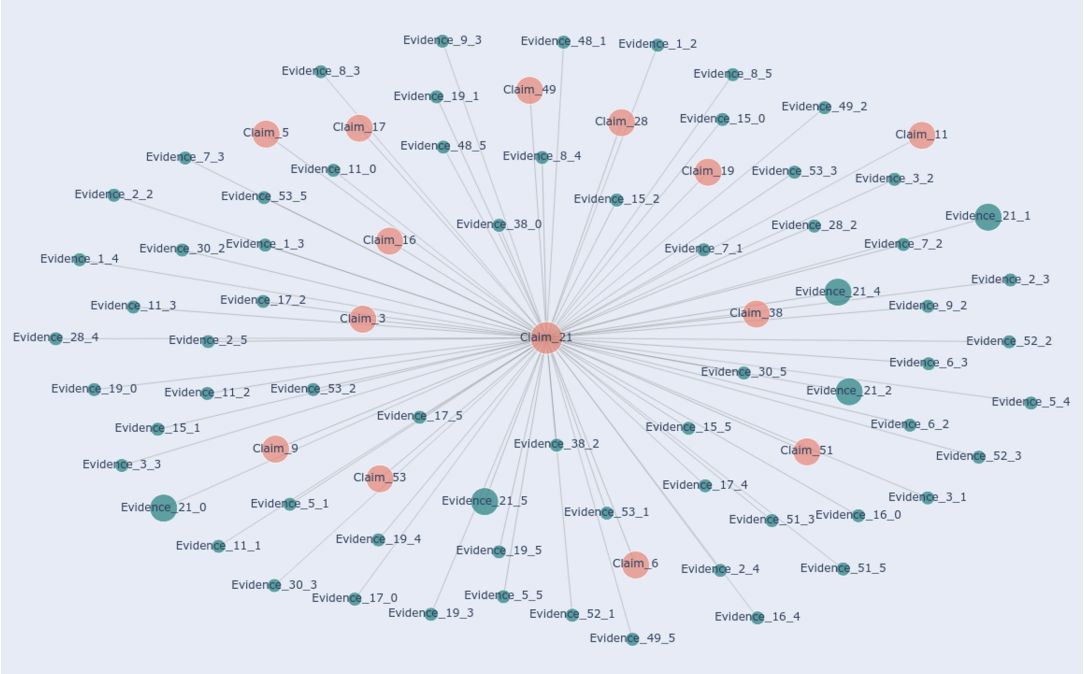


Fig. 5.2 Thematic Interconnections of Claim_21 and its SOI. Highlights Claim_21, its associated evidence and related claims within its SOI. The central claim (Claim_21) is represented by the coral node. Direct evidence of the claim (e.g., Evidence 21_1, Evidence 21_2) are shown as larger teal nodes. Other related claims (e.g., Claim_49, Claim_28, depicted as smaller coral nodes) and additional evidence (e.g., Evidence 48_1, Evidence 1_2, shown as smaller teal nodes) relevant to Claim_21 are included based on our methodology's relevance criteria. Grey lines represent the connections between claims and evidence, showcasing the hierarchical structure and thematic interconnections.

As mentioned in Section 5.3.1, Cluster 0 has 140 samples and their associated evidences from its theme. Before refining the SOI for Claim_21, its interconnections graph reveals 168 nodes and 80 edges, which shows the web of relationships within this specific thematic cluster. The large number of nodes initially represents all possible relationships. After applying a similarity threshold ($\delta = 0.7$ in Equation 5.2), the graph is streamlined to 81 nodes and 80 edges, highlighting the most relevant connections and with more refined and meaningful relationships. This reduction in the number of nodes suggests that many of the initial nodes were either isolated or did not meet the similarity threshold for meaningful connections. The number of edges remaining constant indicates that the relevant relationships were preserved, ensuring that the most pertinent connections between claims and evidences are emphasized. By focusing on these significant relationships, the visualization provides a clear and interpretable representation of both direct associations, such as

the annotated evidence of the claim, and related or indirect associations, including connections with other claims and evidence sentences, within the thematic interconnections of the AFV system.

5.4 Discussion

5.4.1 Implications for Strategic Decision-Making through Integrating Global and Local Perspectives for Comprehensive Explainability

As mentioned earlier, our research efforts aims to enhance explainability in AFV systems by integrating both global and local perspectives. The annotated evidences from our novel dataset, typically comprising six evidence datapoints, which offer a detailed analysis of individual claims, addressing the specific concerns and nuances of each instance, which enables us to compute local post-hoc explainability. Concurrently, our proposed approach in this paper, helps to identify a subset of semantically relevant evidences and claims, offering a macro-level view that highlights broader patterns across diverse factual claims. By combining these perspectives, our methodology seeks to develop a comprehensive framework that not only captures macrolevel thematic patterns but also elucidates micro-level insights into individual instances. This ongoing work is guided by the principles of XAI, focusing on achieving key XAI objectives such as interpretability, accuracy, and fidelity (Ali et al., 2023; Doshi-Velez and Kim, 2017; Guidotti et al., 2018). By adhering to these objectives, we aim to create AFV systems that are both transparent and interpretable, facilitating more reliable and understandable model behaviour.

We also acknowledge that the interim insights derived in this paper, are preliminary and part of an ongoing investigation. Further empirical validation particularly that defined in the Section 5.4.2 and consequent refinement of the methodology are planned for future research.

5.4.2 Future Research Directions

Our future research plans focus on two primary areas: objectives for an upcoming publication and key research directions.

Upcoming Publication Objectives: As mentioned in Section 5.2.1, we will be publishing our novel dataset, which we believe will significantly aid and support

other research efforts in AFV and XAI. The publication will also feature a comparative analysis of in-depth visualization graphs of different claims, highlighting how our methodology does not necessarily incorporate all manually annotated evidences from the dataset. For example, in the case study we examined, Evidence_21_3 was notably absent from the visualization despite being one of the direct pieces of evidence of Claim_21 (see Figure 5.2). This suggests the SOI identifies alternative and potentially more relevant evidence pieces. This insight underscores the value of our approach in providing a deeper and more nuanced understanding of the data, which is critical for high-stakes decision-making scenarios. Additionally, the detailed algorithmic framework of our methodology will also be included in that publication.

Key Research Directions: To build upon the initial work detailed in this paper and further advance our contributions to AFV and XAI, we have outlined a detailed plan for continued research. While the SOI of a claim is currently employed for its visualization, we plan to conduct empirical experiments using the SOI for inference. The envisioned plan is to employ an open-source LLM like Llama2 (Touvron et al., 2023b) to generate post-hoc explanations in two scenarios: with annotated evidences (the six direct evidences from the dataset) and with the aggregate cluster embeddings of the SOI of the selected individual claim; introducing both local and global perspectives to the claim introspection as discussed. This contextual embedding aggregation (Iliadis et al., 2024; Tang et al., 2023; Zhao et al., 2024b) is calculated by averaging the embeddings of all claims and evidences within an SOI of the claim for a cluster. Additionally, we will benchmark our method by conducting a comprehensive performance evaluation using Retrieval Augmented Generation (RAG) (Lewis et al., 2020) to generate explanations for the claim's inference in the context of the entire dataset. This will allow us to assess the effectiveness of our method, as RAG has emerged as a prominent approach in NLP, combining retrieval and generation models to generate more cohesive answers and decrease the occurrence of hallucinations (Salemi and Zamani, 2024). Moreover, the claim-based in-depth graph visualization we proposed in this paper (see Figure 5.2), can serve as a visual aid for global explainability, similar to how LIME (Ribeiro et al., 2016) visualization provides instance-level local explainability. While this holistic approach is expected to provide more comprehensive insights, it requires empirical validation through our ongoing research.

Additionally, we intend to record the results of experiments involving various clustering techniques and similarity thresholds, examining their effects across dif-

ferent themes and clusters. This will allow us to refine our methodology and better understand its robustness and applicability. We also plan to open- source our project on GitHub, ensuring transparency and fostering collaboration by maintaining and version-controlling the code. Furthermore, we will host a prototype space on Hugging Face, enabling practitioners and researchers to interactively test our results and validate the effectiveness of our approach.

5.5 Conclusion

Our ongoing research leverages thematic clustering and contextual visualization to enhance macro- level explainability in AFV systems. By identifying additional relevant evidences, beyond those manually annotated, our methodology enhances decision-making processes through a more comprehensive understanding of the information landscape. This refinement process, is crucial for making informed and accurate decisions, as it highlights the most pertinent connections, providing decision-makers with a clearer view of the evidence base, ultimately supporting more reliable and informed decisions. Guided by XAI objectives, we aim to develop AFV systems that are both transparent and interpretable, facilitating more reliable and deeper understandable model behaviour. As we continue to refine and expand this methodology, we anticipate further insights that will advance the explainability in AFV systems.

Chapter 6

Prelude - Manuscript 3

The advancements in thematic discovery and contextual visualization introduced in Manuscript 2 (Chapter 5) provided key insights into the global structure of AFV. By identifying SOIs and visualizing thematic relationships, it demonstrated the potential of leveraging both local and global perspectives to enhance transparency in AFV. However, while these contributions enabled a deeper understanding of thematic patterns, they were not yet integrated into the retrieval or verification processes of AFV systems.

Manuscript 3 (Chapter 7) builds on these foundations by introducing the CARAG framework, which operationalizes SOI-based thematic embeddings within a retrieval-augmented AFV pipeline. Unlike conventional retrieval systems, CARAG combines claim-specific evidence retrieval with thematic embedding aggregation, ensuring that explanations are not only locally grounded but also globally contextualized. This integration advances explainability in AFV by aligning retrieved evidence with broader thematic structures, enhancing transparency and interpretability.

Another key contribution of Manuscript 3 is the introduction of FactVer, an explanation-focused fact verification dataset curated to support XAI-driven AFV research. This directly addresses the dataset limitations identified in Manuscript 1, where the lack of structured resources for explainability hindered AFV advancements. FactVer provides multi-evidence claims across themes such as COVID, Climate, and Electric Vehicles, offering a benchmark for evaluating both local and global explainability. As with any curated dataset, FactVer's annotation protocol introduces the possibility of bias. Human annotators may emphasize certain thematic framings or overlook minority perspectives, potentially shaping evidence selection in ways that reflect cultural or contextual subjectivity. While this limitation

is acknowledged, later contributions (such as CARAG-u Chapter 9) aim to mitigate such biases by reducing reliance on pre-annotated thematic labels.

CARAG's thematic abstraction through clustering can be contrasted with graph-based retrieval models, which explicitly represent entities and relations through knowledge graphs or graph neural networks. Graph approaches excel at structured reasoning and tracing relational paths, whereas CARAG offers flexibility in open-domain settings without requiring curated graphs. These paradigms are complementary, pointing toward hybrid models that could combine structured graph reasoning with CARAG's adaptive thematic embedding strategy.

This prelude marks the transition from thematic discovery (Manuscript 2) to context-aware retrieval and verification (Manuscript 3). By bridging methodological insights with practical implementation, CARAG advances XAI in AFV, laying the groundwork for further innovations in unsupervised explainability, later explored in Manuscript 4 (Chapter 9).

Chapter 7

CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI (Manuscript 3)

7.1 Introduction

Explainability, defined as the ability to interpret model behavior in a human-understandable way (Doshi-Velez and Kim, 2017), is increasingly essential in AI applications such as Automated Fact Verification (AFV) systems. While recent advancements in AI architectures, such as transformer models (Vaswani et al., 2017) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), have significantly expanded AFV capabilities, they also pose new challenges in ensuring that system decisions remain transparent and interpretable to end-users and decision-makers.

Most AFV systems generally follow a three-stage pipeline architecture, similar to that used in the Fact Extraction and VERification (FEVER) shared task (Thorne et al., 2018b). This architectural approach was later adopted by several subsequent researchers such as Chen et al. (2022b); DeHaven and Scott (2023); Jiang et al. (2021); Krishna et al. (2022); Soleimani et al. (2019); Thorne et al. (2018b); Zhong et al. (2020). This architecture involves the composite tasks of collecting or retrieving relevant evidence to support or refute a claim, ranking these pieces of evidence by impor-

tance, and predicting the claim's veracity, inter alia. However, as AFV systems incorporate more advanced architectures and methods, ensuring interpretability in their increasingly complex decision-making processes has become essential. This need is particularly critical in AFV systems, since online misinformation has become ubiquitous in recent times (Guo et al., 2022). Furthermore, explainability is becoming increasingly important in critical domains such as finance, healthcare, and journalism, and is supported by a growing body of research in Explainable AI (XAI) (Ali et al., 2023; Gunning et al., 2021; Kim, 2018; Vallayil et al., 2023). This emphasis on explainability is also supported by government-led initiatives, such as the European Union's General Data Protection Regulation (GDPR), which mandates explanations for algorithmic decisions, and the United States Department of Defense's (DARPA) XAI program, which aims to make AI systems more interpretable and trustworthy, and highlight the need for greater transparency in AI systems (Ali et al., 2023; Goodman and Flaxman, 2017; Gunning, 2016; Gunning et al., 2021).

However, despite notable advancements in XAI, AFV technologies and the availability of diverse fact verification datasets, the integration of XAI within AFV remains limited. Some studies have attempted to incorporate XAI into AFV systems, although these approaches exhibit certain limitations. For instance, transformer-based models for extractive and abstractive summarization (Atanasova et al., 2020; Kotonya and Toni, 2020b), risk producing incomplete or misleading explanations. Logic-based models like LOREN (Chen et al., 2022a) and ProoFVer (Krishna et al., 2022) create transparent explanations through logic rules but are difficult to scale and apply in real-world contexts. Attention mechanisms (Amjad et al., 2023; Popat et al., 2018; Shu et al., 2019) highlight important features, yet their reliability is questionable as attention scores do not consistently align with key decision-making features (Guo et al., 2022). Counterfactual explanations (Dai et al., 2022; Xu et al., 2023) demonstrate how small input changes affect predictions, but interpreting them remains challenging in complex fact-checking scenarios. Some of these approaches have outperformed the state-of-the-art in claim veracity prediction (Krishna et al., 2022), but they remain limited in terms of scope and effectiveness for explainability.

In addition to the limitations discussed, notable summative studies in this cross-domain of XAI-AFV, such as the comprehensive review on explainable AFV by Vallayil et al. (2023) and the extensive survey on explainable automated fact-checking by Kotonya and Toni (2020a), bring attention to several persistent gaps. Addressing these gaps would not only enhance explainability within AFV systems but also help mitigate the limitations identified in existing approaches. The following items

summarize and discuss some pertinent aspects of explainability in the context of existing approaches.

1. Lack of Explanation-Focused Datasets: Existing fact verification datasets like FEVER (Thorne et al., 2018a) and MultiFC (Augenstein et al., 2019) are not designed to support explanation learning aligned with XAI standards. There is a need for datasets that facilitate training models not only to verify facts, but also to generate meaningful explanations, as previously noted by Stammbach and Ash (2020). This gap in dataset availability limits the development of models capable of both verification and explainability.

2. Overemphasis on Local Explainability: Current explainable AFV systems focus predominantly on local explainability, explaining individual predictions, while neglecting global explainability, which is essential for understanding the system's overall decision-making logic (Kotonya and Toni, 2020a). This local focus leaves AFV systems lacking a holistic view of their behavior, limiting transparency and accountability.

3. Ambiguity in Local and Global Explainability: There is a lack of consensus on how local and global explainability are defined and implemented in AFV systems. While local explainability focuses on individual prediction-level explanations, global explainability refers to understanding the model's overall reasoning process (Zhao et al., 2024a). Different researchers interpret and apply these concepts inconsistently (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Moradi and Samwald, 2021; Murdoch et al., 2019), leading to confusion and retarded progress in explainable AFV research.

4. Inconsistencies in Explainability Taxonomy: There are discrepancies in how explainability concepts are categorized across studies in AFV. For instance, some researchers distinguish between intrinsic and post-hoc explainability (Shu et al., 2019), while others conflate interpretability with explainability, restricting it to individual explanations (Atanasova et al., 2020). This lack of standardization creates further confusion in the field, hindering cohesive advancements in explainable AFV.

In this research, we focus on addressing the first two critical gaps: the lack of explanation-focused datasets and the overemphasis on local explainability. To address these, we propose a comprehensive solution involving a novel explanation-

focused dataset and a context-aware evidence retrieval and explanation generation methodology.

In the subsequent sections, after providing the necessary background in Section 7.2, we describe the dataset in Section 7.3 and the methodology in Section 7.4. The dataset introduced is curated for XAI research in AFV. It pairs each claim with multiple annotated pieces of evidence within its thematic context (e.g., Climate change, COVID-19, Electric Vehicles). The dataset facilitates both local and global explainability by enabling deeper exploration of claim-evidence relationships and thematic patterns extending beyond individual data points, while also supporting explainability-focused studies. Meanwhile, our context-aware retrieval methodology enhances the AFV pipeline, particularly the retrieval component, by incorporating thematic embeddings generated from a subset of the fact verification dataset. This subset is identified through a statistical modeling approach and further refined through a semantic aggregation technique. By integrating broader contextual information with claim-specific embeddings, this methodology not only advances existing frameworks like RAG but also introduces a broader thematic context, resulting in more nuanced and context-sensitive explanations. The experimental framework of the methodology, including a case study and comparative analysis with RAG, is presented in Section 7.5, while challenges, future research directions, and conclusions are discussed in Sections 7.6, 7.7, and 7.8, respectively.

While our work contributes to mitigating these issues, the remaining two challenges, ambiguity in explainability terminologies and inconsistencies in explainability taxonomy, are expected to be gradually refined as more research in the field emerges, leading to greater clarity and standardization.

7.2 Background

The evolution of AFV systems began with 'knowledge-free' approaches relying solely on linguistic features of claims for verification, without using external evidence (Rashkin et al., 2017). This was followed by the integration of structured knowledge bases, like RDF triples, for fact verification, but faced challenges with scalability and handling nuanced information (Gardner and Mitchell, 2015; Shi and Weninger, 2016). A major advancement in AFV systems came with the introduction of evidence retrieval methods, where claims were verified against retrieved textual sources, such as Wikipedia, as demonstrated in the FEVER dataset by Thorne et al. (2018b). While Wikipedia offered broad accessibility, it also introduced challenges

with comprehensiveness and potential biases, impacting the fidelity of resulting AFV models (Hanselowski et al., 2019). A further breakthrough was achieved with the development of advanced retrieval capabilities, exemplified by RAG, which dynamically integrate external knowledge during inference to enable more context-sensitive and informed veracity predictions (Singhal et al., 2024). Nonetheless, interpretability, accuracy, and fidelity remain essential paradigms in explainable AI, as emphasized by recent work on XAI (Ali et al., 2023; Doshi-Velez and Kim, 2017; Guidotti et al., 2018).

Maintaining these XAI principles has become increasingly challenging due to the evolving complexity of modern AFV systems, particularly with the use of pre-trained Foundation Models (FM) and Large Language Models (LLM) in different roles across the AFV pipeline. For instance, LLMs are used as encoders for embedding generation to capture semantic representations of claims and evidence, for natural language inference (NLI) in veracity prediction, and for natural language generation (NLG) in crafting coherent explanations. Additionally, incorporating RAG for dynamic evidence retrieval and in-context learning for veracity assessment based on retrieved evidence (Singhal et al., 2024) has further increased the complexity of these systems. In particular, the use of LLMs, with their massive scale in terms of parameters and training data, presents unique challenges for explainability in downstream tasks like AFV. Moreover, these models require extensive computational resources for generating explanations. Consequently, established interpretability methods, including feature attribution methods such as gradient-based approaches (Sundararajan et al., 2017) and SHAP values (Lundberg and Lee, 2017), as well as surrogate models like LIME (Ribeiro et al., 2016), can become computationally impractical for explaining models with billions of parameters, limiting their feasibility for current AFV systems compared to traditional deep learning models.

In addition to the influence of the operational role of LLM integration we discussed, the training paradigm adopted for the employed LLM also necessitates diverse strategies for achieving XAI in AFV. Specifically, the approach to explainability varies significantly based on whether the model is fine-tuned or used directly through prompting. In the fine-tuning paradigm, models like BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), which are pre-trained on large corpora, are subsequently fine-tuned on labeled datasets for specific tasks, such as AFV. In contrast, the prompting paradigm utilizes models without additional training, as seen with base models like GPT-3 (Brown et al., 2020) and Llama 3 (Dubey et al., 2024), which respond based on pre-trained knowledge; or as with assistant models like GPT-4

by OpenAI (Bubeck et al., 2023), Claude by Anthropic (AnthropicAI, 2023) which undergo additional alignment through methods like instruction tuning and Reinforcement Learning from Human Feedback (RLHF) to perform user-specific tasks (Zhao et al., 2024a). The prompting paradigm is further reinforced by the impressive zero-shot performance of LLMs in various language tasks (Liusie et al., 2024), showcasing their capability to handle complex tasks without task-specific fine-tuning. These diverse methods of employing LLMs significantly affect how XAI research in modern AFV is approached. Fine-tuned models require tailored interpretability methods that account for task-specific adjustments, whereas prompting-based models rely on post-hoc explanations generated from the models' pre-trained knowledge (Zhao et al., 2024a).

Furthermore, the intended scope of explainability, whether local or global, further influences the choice of XAI strategies devised for AFV. Therefore, effective XAI in AFV must consider the model's operational role (e.g., encoder, NLI, NLG), the training paradigm employed (fine-tuning, few-shot, or prompting), and desired explainability scope (local or global), necessitating a holistic approach. However, as outlined in the introduction, current XAI methods in AFV primarily involve post-hoc explanations (i.e., methods applied after the model has been trained to explain its predictions), including transformer-based models (extractive and abstractive summarization to assist veracity prediction), logic-based models (using logic rules to create transparent explanations), attention mechanisms (highlighting important features), and counterfactual explanations (showing how small input changes affect predictions), each with limitations in scalability, reliability, and interpretability.

In this research, we advance post-hoc explanations by enhancing both the retrieval and generation components of RAG: incorporating thematic embeddings for context-aware evidence retrieval and leveraging zero-shot NLG with optimized LLM prompting for abstractive summarization. Section 7.4 details our framework, addressing the roles, paradigms, and scope of XAI in AFV comprehensively for a balanced explainability.

Table 7.1 Key Components of FactVer

Header	Description
Claim_Topic_ID	A unique identifier representing the claim, which also encodes information about the thematic topic it belongs to. For example, in the identifier Claims_Climate_B2.0_1, the middle segment ('Climate') denotes the theme, while B2.0 refers to the annotation team responsible for curating the claim, facilitating efficient data processing.
Claim_Text	The textual content of the claim, which needs to be verified for its truthfulness.
Evidence_Topic_ID	A unique identifier for the evidence corresponding to each claim. The Evidence ID is constructed by consolidating the claim ID with the prefix 'Evidence' and appending a unique serial number. For example, in Evidence_Claims_Climate_B2.0_1_n, the 'n' uniquely distinguishes each piece of evidence associated with the claim. This structure efficiently organizes the multiple pieces of evidence for a given claim.
Evidence_Text	The actual textual evidence supporting or refuting the claim.
Label	The label indicating the veracity of the claim (T/F/N representing True/False/Not Enough Info respectively).
Reason	An explanation that provides justification for the label assigned to the claim.
Reason_Type	Classifies the nature of the explanation as either Abstractive for human-generated explanations that are crafted based on the evidence but not directly copied, or Extractive, copied directly from the supporting evidence. If no explanation is provided, it is marked as Nil.
Annotation_ID	An identifier assigned to each entry, reflecting the annotation team responsible for curating the data. There are three types of IDs, *B_2.0*, *C_2.1*, *C_2.2*, corresponding to the three annotation teams involved, allowing for traceability back to the raw files from each team during data processing and analysis.
Article_Topic_ID	A reference to the specific source or article from which the evidence is derived. This ensures the data can be linked back to the original source used by the teams during the annotation process.

7.3 Dataset

In this section, we introduce ***FactVer***, a novel dataset developed to address key limitations in existing AFV datasets by supporting both fact verification and XAI research, with a focus on enhanced transparency and explainability. Aligned with recent research directions in explainable AFV, such as those proposed by Vallayil et al. (2023), the dataset offers structured evidence relationships and human-generated explanations across multiple topics. By enabling deep exploration of claim-evidence relationships and thematic patterns, the dataset facilitates both local and global explainability, paving the way for advanced research in explainability-focused AFV systems.

7.3.1 Structure and Composition

The dataset is organized into the following thematic topics and structured around key components corresponding to its column headers, as outlined in Table 7.1.

- *Climate Change*: Claims and evidence related to global warming, environmental policies, and their socioeconomic impacts.
- *Covid-19*: Claims and evidence concerning the pandemic, vaccines, treatments, and public health measures.
- *Electric Vehicles*: Claims and evidence focused on electric vehicle technology, battery innovations, efficiency, and market trends.

The dataset was generated through a rigorous annotation process, ensuring consistency across themes while capturing diverse perspectives. The following sections provide detailed statistics about the dataset, describe the preparation process, and present example data entries to illustrate its structure and composition.

7.3.2 Dataset Description

The dataset includes 589 unique claims. As shown in Figure 7.1, the majority of these claims are supported by 6 pieces of evidence (approximately 70%), while a smaller subset of claims has only 1 piece of evidence (about 22%). A minor portion of claims is associated with 12 pieces of evidence (around 7%). This range of evidence distribution provides flexibility in terms of the depth and complexity of explainability within the fact verification process.

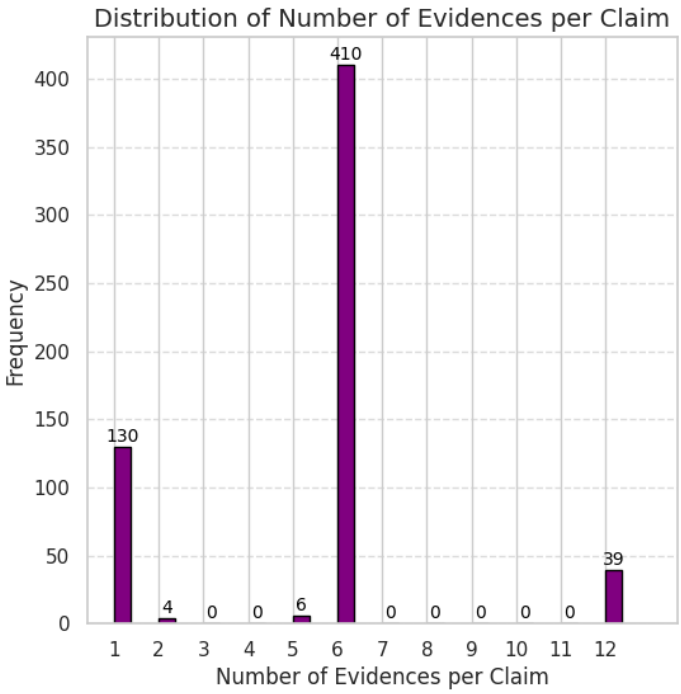


Fig. 7.1 Distribution of the Number of Evidences Associated with each Claim Across All Themes. The figure illustrates that most claims are supported by six pieces of evidence(represented by the tallest bar), with smaller subsets having either one or twelve pieces of evidence, highlighting the variability in evidence distribution across the dataset.

Furthermore, Figure 7.2 illustrates the distribution of text length for *Claim_Text*, *Evidence_Text and Reason*. The following key points can be observed:

- *Claim_Text Length*: The majority of claims are concise, typically within the 40-60 word range. Longer claims exceeding 100 words are less common.
- *Evidence_Text Length*: Evidence text length varies widely, with most evidence ranging from 150-250 words, though some extend up to 700 words, reflecting varying levels of detail required to support different claims.
- *Reason Length*: Reasons are generally concise, with most falling within the 50-100 word range. While some explanations exceed 200 words, very few extend beyond 400 words.

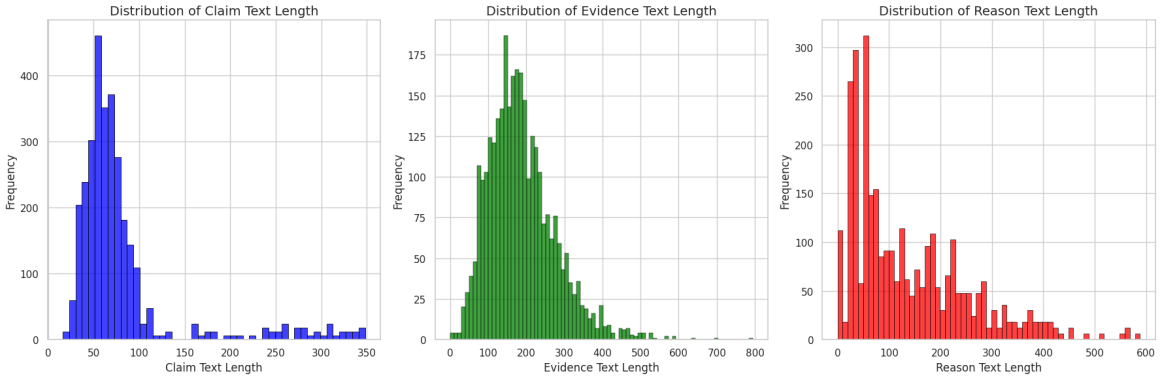


Fig. 7.2 Distribution of Claim, Evidence, and Reason Text Length

This distribution highlights the diverse nature of the dataset, where claims, evidence texts, and human-generated reasons exhibit varying lengths, potentially reflecting the differing complexity of fact verification instances.

7.3.3 Preparation

The dataset was constructed from a large corpus of news articles, sourced from over 20,000 searches conducted via Google, Bing, and DuckDuckGo between October and December 2022 across the three themes/topics mentioned in Section 7.3.1, collectively referred to as ***FactVer_1.0***. These articles were processed to extract key information, including titles, body content, and relevant metadata (e.g., publication date, article ID, and URL), resulting in the intermediate version named ***FactVer_1.1***.

Following this, three independent annotation teams worked with the processed dataset using a unified instructions document. Annotators followed step-by-step instructions in that document for creating a fact verification dataset for their assigned topics, generating claims based on the article content and identifying corresponding evidence spans. It is important to note that each annotation team worked on non-overlapping topics (as specified in the instructions), annotating separate subsets of the dataset, and inter-annotation agreement was hence not applicable. Each claim received a unique *Claim ID*, and evidence pieces were labeled with unique *Evidence IDs* (e.g., E1 to E6) for traceability. Claims were labeled as True (T), False (F), or Not Enough Info (N) based on the evidence provided. Annotators also included a *Reason* field, offering explanations for the assigned labels, which could either be derived directly from the evidence or be a novel, human-generated explanation.

This process resulted in three intermediate fact verification datasets, collectively referred to as ***FactVer_1.2_X***, where X represents the *Annotation_ID* of each respective

team. Further details about the annotation process, including the template provided to annotators and the instructions they followed, are available in Appendix A.1. Although the annotation guidelines recommended supporting each claim with up to six pieces of evidence, the actual number of evidence pieces per claim in the consolidated fact verification dataset ranges from 1 to 12, as discussed in Section 7.3.2 and represented in Figure 7.1, reflecting the varied interpretations and approaches of the annotation teams.

Building on these intermediate datasets, their consolidation resulted in a unified dataset created through additional data cleaning, preprocessing, and traceability steps (details of which are also provided in Appendix A.1). This consolidated dataset, named ***FactVer_1.3***, is designed to facilitate AFV and support XAI research in this domain.

7.3.4 Example Data Entries

To illustrate the dataset structure, we provide examples of data entries, using the first claim (Claims_Climate_B2.0_1) in the dataset as a representative example.

- *Claim_Topic_ID: Claims_Climate_B2.0_1*
- *Claim_Text: New Zealand has a carbon trading system*
- *Label: T*
- *Evidence_Text: (The following list contains the six pieces of annotated evidence associated with this claim:)*
  - *Evidence 1: A number of other countries have, however, also implemented a carbon trading system at a national or sub-national level, or have one in development, including Canada, China, Japan, New Zealand, South Korea, Switzerland, and the United States, according to the European Commission.*
  - *Evidence 2: As of July, 46 countries are pricing emissions through carbon taxes or emissions trading schemes (ETS), according to the International Monetary Fund.*
  - *Evidence 3: NZ's agricultural emissions aren't currently captured under the ETS (unlike other sources like industrial processes).*
  - *Evidence 4: The number of emission units released for auction is designed to meet New Zealand's international obligations.*

  - *Evidence 5: With many New Zealand farms having been converted to forestry due to rising carbon prices in recent years, feedback last week closed on a proposal to change settings in the Emissions Trading Scheme, where permanent plantings of exotic forests, like pine forests, would be excluded from the scheme from next year.*
  - *Evidence 6: China, South Korea, Canada, Japan, New Zealand, Switzerland and the US already have a number of national or regional systems, however the international carbon market is said to develop through a bottom-up approach, whereby the EU ETS will be linked with other international systems, with a common aim to reduce the amount of emissions.*

- *Reason: New Zealand has an existing emissions trade scheme*
- *Reason_Type: Abstractive*

7.3.5 Summary

FactVer, designed to advance explainability-focused research, addresses the need for datasets that support both fact verification and explanation learning. Its structured evidence annotations and diverse thematic scope provide a valuable resource for improving fact-checking methods and advancing AI-driven research in both local and global explainability within AFV systems. The dataset is integral to our methodology, providing a foundation for developing and validating new approaches in AFV, as discussed in the following section.

To ensure reproducibility and foster further research, the dataset is publicly available on Hugging Face[1], and the associated code is available on GitHub[2].

7.4 Methodology

This section presents our Context-Aware 'Retrieval Augmented Generation' Framework (CARAG), an approach to enhance evidence retrieval and post-hoc explanation generation in AFV systems. Traditional retrieval methods often process each query in isolation, overlooking the broader (or non-local) context surrounding a claim. CARAG addresses this gap, leveraging the structured evidence and

[1]Dataset available at: https://huggingface.co/datasets/manjuvallayil/factver_master

[2]Code repository available at: https://github.com/manjuvallayil/factver_dev

thematic insights from the ***FactVer*** dataset, ensuring that retrieval aligns with both claim-specific details and its broader thematic background, leading to more informed prompts for LLMs and, consequently, richer fact verification explanations, as elaborated in the following sections.

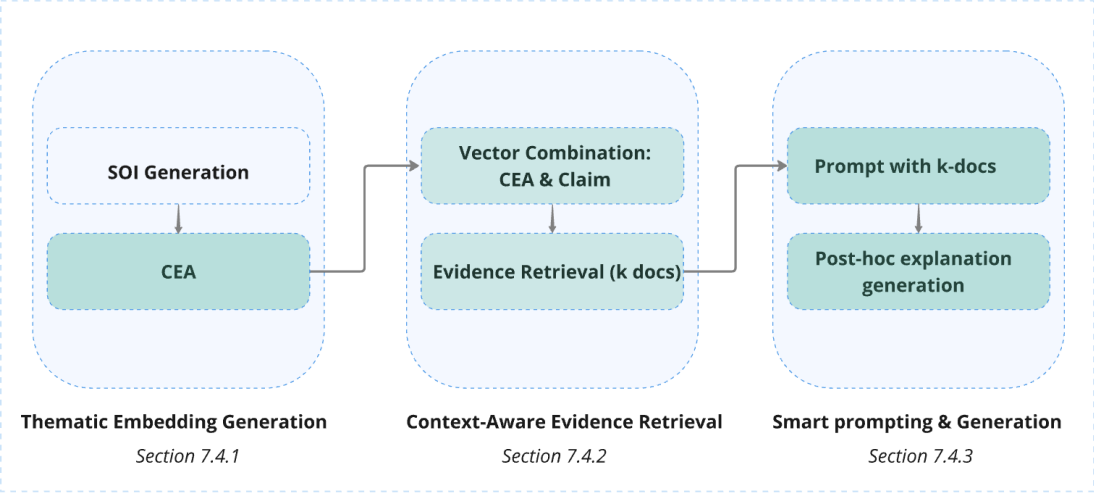


Fig. 7.3 Overview of the Methodology Components

Figure 7.3 presents a visual overview of our methodology, summarizing the key components described in the subsequent sub-sections in methodology description. It simplifies the understanding of our otherwise intricate process by offering a step-by-step representation of how different phases, retrieval and generation, interact. Later in this section, a more detailed diagram showcases how CARAG is integrated into the complete AFV pipeline, demonstrating how it refines both evidence retrieval and explanation generation compared to standard methods.

7.4.1 Thematic Embedding Generation

The first step in our methodology is to generate thematic embeddings by leveraging the Subset of Interest (SOI), a concept introduced to find non-local context of a claim under investigation (Vallayil et al., 2024). In that work, the SOI was utilized for cluster visualization, offering insights into both the claim's annotated evidence and also its broader, non-local context. The SOI generation process starts by identifying the theme of the selected claim (e.g., climate change) from the fact verification dataset. The dataset is then filtered to retain only claims and evidence relevant to the identified theme. This thematic subset is then structured through

clustering, organizing semantically similar claims and evidence into distinct groups based on their embeddings, as further elaborated in Section 7.5.1.

The cluster containing the selected claim is then identified, and all items within this cluster are extracted to form an initial set. This set includes: (i) the selected claim, (ii) its directly annotated evidence (if any in the same cluster), (iii) other claims within the identified cluster (hereafter referred to as *related claims*), and (iv) the annotated evidence of these related claims (if available within the same cluster, and here after referred to as *thematic cluster evidence*). Importantly, not all annotated evidence of the selected claim or related claim's will necessarily be present in the identified cluster, as it is the result of an unsupervised clustering process.

Table 7.2 Key Components of the SOI Dictionary

Key	Description
claim_id	The unique identifier of the selected claim.
claim	The text of the selected claim.
annotated_evidences	The evidence pieces directly associated with the claim, including both the text and unique evidence IDs.
related_claims	Other claims within the thematic cluster that are semantically related to the selected claim.
thematic_cluster_evidences	Evidence from other claims within the thematic cluster that are thematically relevant to the selected claim, including both text and unique evidence IDs.
similarities	The similarity scores between the selected claim and the associated evidence/related claims, calculated using cosine similarity.

Next, cosine similarity is calculated individually between the embedding of the selected claim and the embedding of each item in this initial set. Items that do not meet the empirically chosen similarity threshold ($\delta = 0.75$) are excluded. This threshold was selected as a balanced criterion to filter out loosely related instances while retaining a thematically relevant subset of the fact verification corpus for a given claim. The resulting refined subset forms the SOI of the selected claim. The SOI is stored in a dictionary format, as shown in Table 7.2, containing the claim details, its directly annotated evidence pieces, related claims, thematic cluster evidence, and cosine similarity scores quantifying the relevance of each evidence or claim to the selected claim. (The complete algorithm for SOI generation is provided in the Appendix A.2 for those interested in the technical details).

However, while the possibility of leveraging the SOI of a claim for evidence retrieval or inference mechanisms within the AFV pipeline was highlighted as a future direction in prior work (Vallayil et al., 2024), it was not implemented. In this research, we extend the utility of SOI by integrating it into the AFV pipeline for the first time, moving beyond its visualization purpose. This transformation evolves the SOI from a static visual representation into an active component of CARAG.

To implement this SOI integration into CARAG, selected elements from the SOI dictionary are extracted, specifically, the SOI['annotated_evidences'], SOI['related_claims'], and SOI['thematic_cluster _evidences']. Each item in these fields is then passed through a Sentence Transformer model (SBERT, specifically the all-mpnet-base-v2 variant), which converts the text into a numerical vector (embedding) that captures the semantic meaning and contextual relevance of the text.

$$\text{Thematic\_Embedding} = \frac{1}{n}\sum_{i=1}^{n}\text{Embedding}(e_i), \quad e_i \in \begin{cases} \text{SOI['annotated\_evidences']}, \\ \text{SOI['related\_claims']}, \\ \text{SOI['thematic\_cluster\_evidences']} \end{cases} \tag{7.1}$$

Next, as represented in Equation 7.1, a single unified thematic embedding is generated by calculating the element-wise average of embeddings corresponding to these specific elements within the SOI. This involves summing the element wise numerical components of the embeddings and dividing the result by the total number of embeddings. Embedding aggregation techniques, such as averaging, or graph-based methods, have been explored in various studies (Iliadis et al., 2024; Tang et al., 2023; Zhao et al., 2024b).

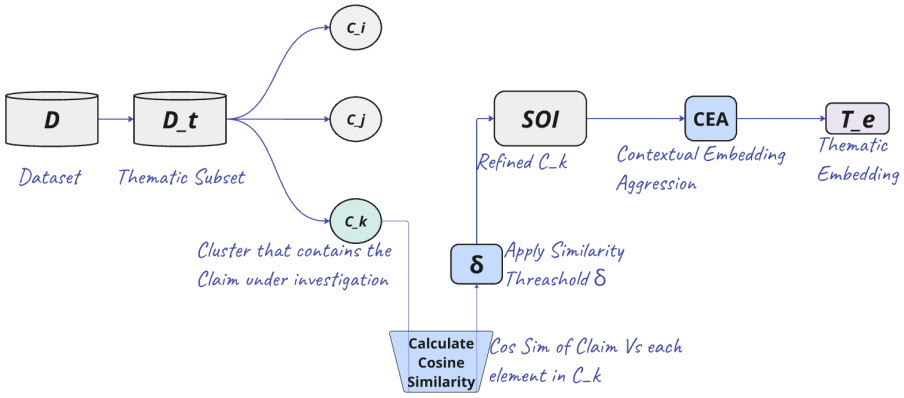


Fig. 7.4 The figure illustrates the curation of the unified thematic embedding (T_e) process from the fact verification dataset (D). The Claim's cluster (C_k), identified from the thematic subset (D_t), is refined using cosine similarity and a similarity threshold (δ) to form the Subset of Interest (SOI) for the claim. This SOI is then processed through Contextual Embedding Aggregation (CEA) using Equation 7.1 to generate the final thematic embedding.

In our pipeline, we define this phase as Contextual Embedding Aggregation (CEA), and Figure 7.4 illustrates the process of generating the thematic embedding (represented as (T_e), starting from the dataset (D). This aggregated thematic embedding, derived from the SOI and capturing both local and global contexts, serves as the foundation for the explainable AFV framework we propose in this study, particularly for the evidence retrieval process.

More importantly, while this thematic embedding encapsulates the broader context of the claim derived from the SOI dictionary, the claim itself is excluded from the CEA process. As represented in Equation 7.1, the claim embedding is deliberately omitted from this computation. This distinction ensures a clear separation between the claim-specific embedding and the contextual embedding, which are later combined during the retrieval process (detailed in Section 7.4.2).

7.4.2 Context-Aware Evidence Retrieval

Building on the thematic embeddings generated through CEA, this stage of CARAG integrates them into the evidence retrieval process by combining the claim vectors with the thematic embedding, which together serve as the query for retriev-

ing evidence from the (vectorized) fact verification database. These embeddings are merged using a weighted mechanism (Equation 7.2), where the parameter α controls the balance between claim-specific details and thematic context.

$$\text{Combined\_Embedding} = \alpha \cdot \text{Claim\_Embedding} + (1 - \alpha) \cdot \text{Thematic\_Embedding} \tag{7.2}$$

As established in Section 7.4.1, the claim text is not included in the CEA process, ensuring that it remains distinct and is combined separately with the thematic embedding during this stage. This separation supports flexible weighting, allowing for varying influences of claim-specific and contextual details depending on the task, and can extend to other retrieval tasks requiring a balance between localized and contextual information. An α of 0.5 aims to achieve an equal balance of claim-specific details and broader contextual information in the retrieval query, thereby influencing the selection of evidence with contextual insights.

7.4.3 Smart Prompting for Explanation Generation

Following the context-aware evidence retrieval process, this stage introduces natural language generation into the framework, where an LLM is employed to generate concise fact-verification explanations. The retrieved evidence is incorporated into the LLM prompt alongside the claim text and specific instructions. By ensuring that the evidence is enriched with both specific and contextual insights, the prompt is crafted to reflect a more comprehensive perspective, integrating information that goes beyond the immediate claim. This results in more human-readable explanations that are both informative as well as contextually grounded, as we will further discuss in subsequent sections.

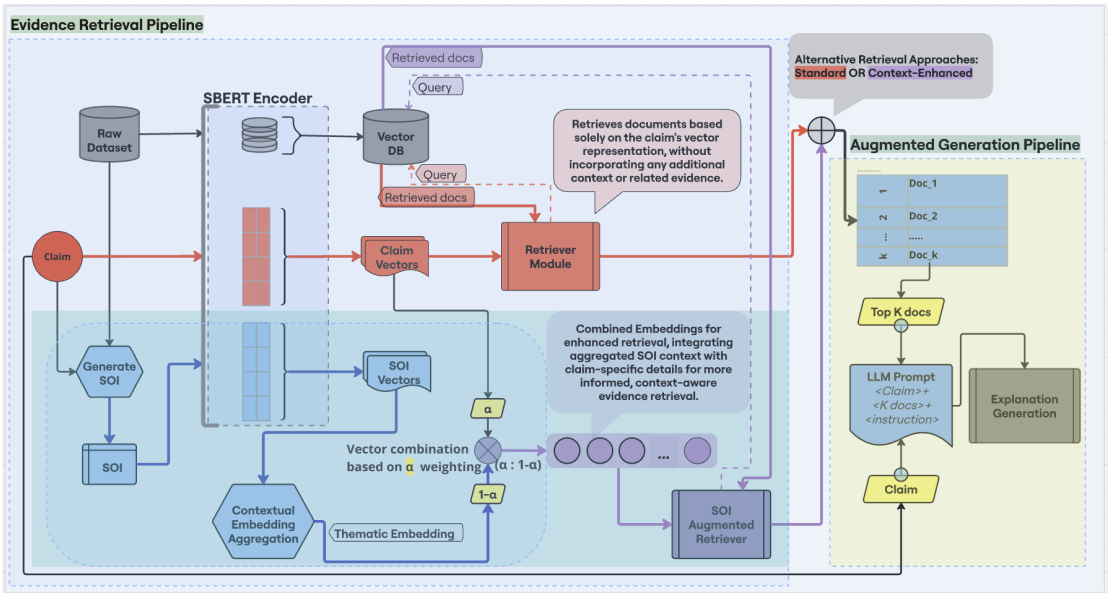


Fig. 7.5 Overview of the CARAG Framework vs. Standard RAG. The Standard Retrieval path (red arrows) retrieves evidence solely based on the claim vector, without incorporating contextual information. In contrast, the CARAG pipeline (light blue-shaded area) introduces context-aware evidence retrieval, which includes the generation of SOI, followed by thematic embedding aggregation and its weighted combination with claim embeddings for evidence retrieval. Retrieved documents are then passed to the Augmented Generation Pipeline (yellow-shaded area), where an LLM generates explanations based on the claim, retrieved evidence, and an instructional prompt.

Figure 7.5 provides a visual comparison between the standard AFV pipeline and the CARAG framework, while also summarizing our methodology. The Standard Retrieval path (red arrows) follows a conventional approach, retrieving evidence solely based on the claim vector, without accounting for any contextual information. In contrast, the CARAG framework (light-blue-shaded area) generates the thematic embedding (blue arrows) as described in Equation 7.1, and combines it with the claim vector using a weighted averaging process (Equation 7.2). This weighted approach produces a more refined final combined embedding used for querying (represented by the purple dots in Figure 7.5), offering a more nuanced integration compared to simple vector concatenation. The retrieved documents are then passed to the Augmented Generation Pipeline (yellow-shaded area), where an LLM prompt is constructed by combining the claim, the top-k retrieved evidence, and specific instructions. The LLM (e.g., LLaMA) subsequently generates a concise explanation, assessing the claim's veracity and offering a justification.

7.5 Experimental Framework & Results

This section describes the key elements employed in our experimental framework and outlines a case study comparison of explanation approaches (Section 7.5.1) alongside a comparative analysis of RAG and CARAG methods (Section 7.5.2). The experimental framework integrates custom Python modules for data management, clustering, embedding generation, and fact verification, leveraging purpose-built methods and pre-trained models for embedding and explanation generation.

For embedding generation, we selected the Sentence-BERT (SBERT) model (Reimers and Gurevych, 2019), which enhances BERT by incorporating siamese and triplet network structures to produce semantically meaningful sentence vectors. Specifically, we employed the open-source *all-mpnet-base-v2* variant of SBERT, fine-tuned on over 1 billion textual pairs. This model is relevant in our methodology, where cosine similarity supports context filtering and nuanced textual similarity (Jayanthi et al., 2021). Moreover, we chose to use the same SBERT encoder as employed in the SOI methodology (Vallayil et al., 2024) to ensure consistency and comparability. Exploring alternative embedding models to assess their impact on the pipeline's performance remains an avenue for future work.

For evidence retrieval tasks, we integrated the FAISS (Facebook AI Similarity Search) library (Douze et al., 2024) into our pipeline. FAISS enables rapid similarity searches on large datasets, managing vectorized storage of our corpus to facilitate document retrieval. By indexing the *all-mpnet-base-v2* embeddings generated from our dataset, FAISS scales evidence retrieval efficiently. This setup allows both RAG and CARAG to retrieve evidence directly from the indexed vectors, thereby supporting explanation generation and subsequent processes.

For fact verification and explanation generation, we employed the *Llama-2-7b-chat-hf* variant of LLaMA from Meta (Touvron et al., 2023a), chosen for its balance of efficiency and performance and its compatibility with our computational resources. With 7 billion parameters, Llama-2 Chat is suited to our explainability tasks, offering competitive performance comparable to models like ChatGPT and PaLM (Touvron et al., 2023a). Optimized for dialogue and trained with RLHF, the model supports our informed prompting methodology (Section 7.4.3) to generate coherent, user-aligned explanations. Implemented using the *AutoTokenizer* and *AutoModelForCausalLM* classes from Hugging Face, the process follows a sequence-to-sequence (seq-to-seq) approach, where the input sequence combines the claim text, retrieved evidence, and an instructional prompt. The output sequence includes natural language reasoning,

providing a verdict on the claim and a nuanced post-hoc explanation. Operating in zero-shot mode, the model leverages its pre-trained linguistic and contextual capabilities without task-specific fine-tuning. Example of this workflow is detailed in Section 7.5.1. To ensure unbiased results, GPU memory is cleared before each generation run. Future work could explore newer versions, such as Llama 3 (Dubey et al., 2024), on advanced hardware to assess potential improvements.

To visualize the thematic clusters and SOI of a claim, as shown in Figure 7.6, we used the NetworkX Python package and Plotly's graph objects. Additionally, for the comparative visual analysis of RAG and CARAG, we rely on scikit-learn and seaborn to apply PCA (Principal Component Analysis), t-SNE (t-distributed Stochastic Neighbor Embedding), and KDE (Kernel Density Estimation). These techniques helped us to simplify high-dimensional embeddings, preserve both local and global patterns, and generate thematic density contours, as detailed in Section 7.5.2.

This architecture enables streamlined integration of the AFV pipeline, supporting both RAG and CARAG methods for a comprehensive analysis, as shown in Figure 7.5 and empirically discussed in subsequent sections as we evaluate the framework using instances from our fact-verification dataset *FactVer*, introduced in Section 7.3.

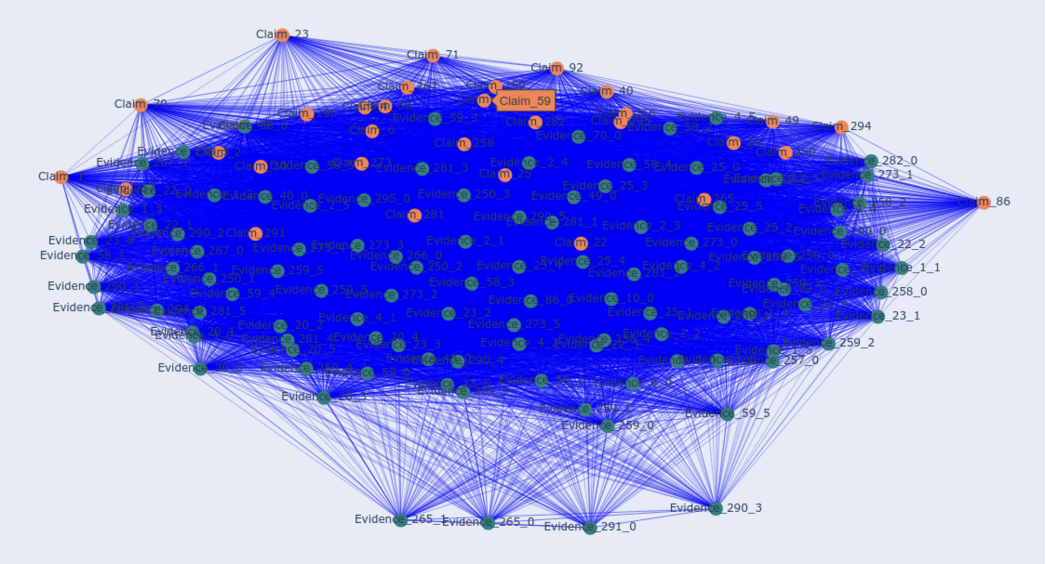
7.5.1 Case Study Analysis of CARAG

In this section, we present a focused case study analysis to illustrate an end-to-end experimental evaluation of our framework. For this, we selected the claim, *"The public is unconcerned about a climate emergency"* (Claim 59) from *FactVer*. This claim serves as a representative example, allowing us to illustrate CARAG's performance on a complex, real-world issue. Additionally, Claim 59 was chosen due to its nuanced nature; the human-generated (abstractive) explanation for this claim in *FactVer* is, "There is not enough evidence to suggest that people are concerned or unconcerned with the climate emergency." This highlights the ambiguity and contextual depth required in handling such claims, making it an ideal test case for evaluating CARAG's capabilities.

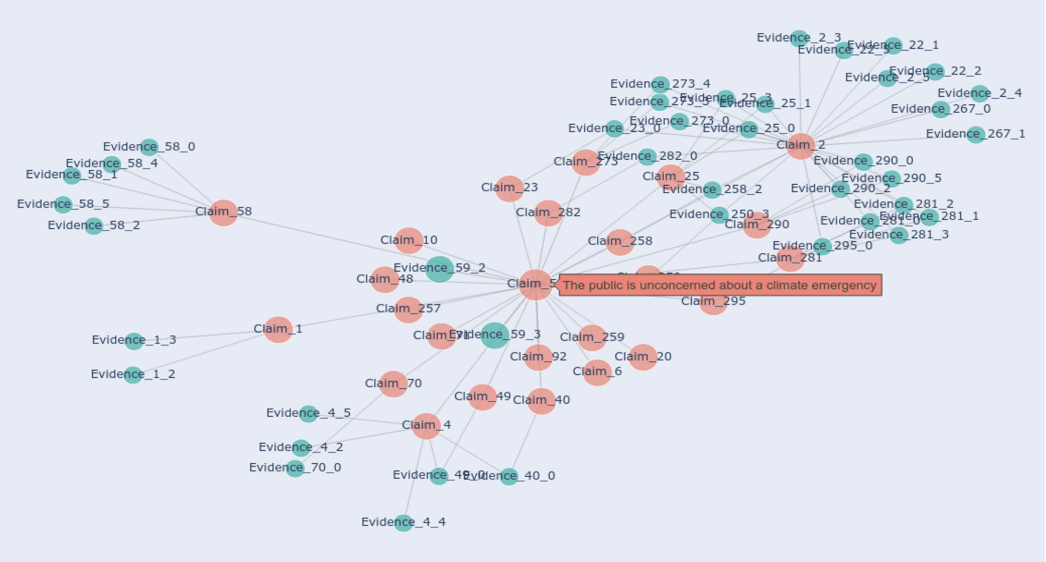
The case study description follows the exact procedural order of our methodology, with the sub sections below corresponding to Sections 7.4.1, 7.4.2, and 7.4.3, respectively, illustrating the practical application of our structured methodology for Claim 59.

Thematic Embedding Generation for Claim 59

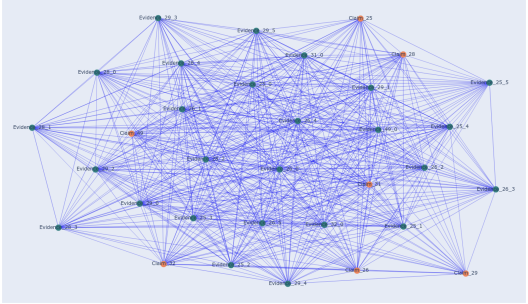
To generate a thematic embedding for Claim 59 from *FactVer* using CEA (Section 7.4.1), we first needed to identify a focused subset of contextually relevant data that would form the basis of our analysis. This involved applying our SOI approach to determine the theme associated with Claim 59 (Climate), then filtering the corpus to retain only instances within this theme, ensuring alignment with the claim's context. As outlined in Section 7.4.1, this thematic subset is then structured through clustering to organize semantically similar claims and evidence into distinct groups. To achieve this, we applied GMM-EM clustering (Al-Dujaili Al-Khazraji and Ebrahimi-Moghadam, 2024; Jiao et al., 2023) within the Climate theme, identifying three unique clusters: Cluster 0, Cluster 1, and Cluster 2. The selection of GMM-EM is motivated by its effectiveness in identifying underlying patterns in complex data, with prior applications in speaker identification, emotion recognition, and brain image segmentation (Al-Dujaili Al-Khazraji and Ebrahimi-Moghadam, 2024; Jiao et al., 2023). In this context, we adapt it to model the dataset as a combination of multiple thematic structures, capturing structural similarities between claims and evidence. Our methodology employs GMM in a hard clustering approach, assigning each claim and evidence to a single cluster to ensure clear relationships and facilitate precise analysis in AFV (Vallayil et al., 2024). Claim 59 is identified within Cluster 1, a dense network containing 85 nodes and 3,103 edges, indicating a rich interconnection of semantically related claims and evidence. Following the methodology outlined in Section 7.4.1, we refined this cluster using a cosine similarity threshold of $\delta = 0.75$ to retain thematically relevant claims and evidence. The resulting SOI dictionary for Claim 59 incorporates all the fields presented in Table 7.2, providing a structured foundation for embedding generation.



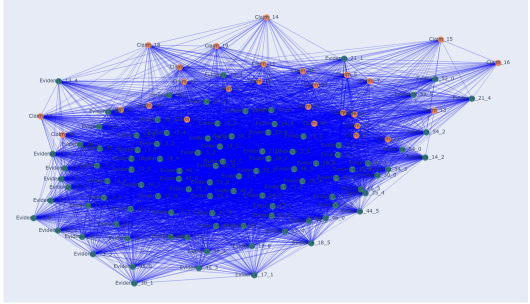
(a) Cluster 1: Full thematic cluster containing Claim 59



(b) SOI for Claim 59: Subset Derived from Cluster 1



(c) Cluster 0: Secondary thematic cluster for comparison



(d) Cluster 2: Tertiary thematic cluster for comparison

Fig. 7.6 Visualization of SOI and thematic clusters within the Climate theme. Panels (a), (c), and (d) depict the identified clusters (Clusters 1, 0, and 2, respectively). In Panel (a), Cluster 1 highlights Claim 59 for clarity, while Panel (b) shows the refined SOI for Claim 59 derived from Cluster 1.

Figure 7.6 provides a visualization of the thematic clusters for Claim 59. Panels (a), (c), and (d) display the three distinct clusters identified through GMM-EM clustering, Cluster 1, Cluster 0, and Cluster 2 respectively, illustrating thematic separation within the Climate theme. Cluster 1, shown in panel (a), is of particular interest as it contains Claim 59 along with the most thematically relevant connections for our analysis. For this reason, we present Cluster 1 alongside its refined SOI, derived from this cluster, as bigger sub plots (panels a and b), allowing for a direct comparison between the full thematic cluster(Cluster 1) and its distilled subset(SOI). Compared to Cluster 1, Cluster 0 (panel c) is more sparsely connected, whereas Cluster 2 (panel d) is denser. This variation in density underscores the GMM-EM algorithm's flexibility in clustering, as it naturally groups conceptually related data based on thematic relevance rather than enforcing uniform cluster sizes. This approach ensures that each cluster accurately reflects the underlying thematic nuances within the broader climate context.

In the SOI graph in panel (b), Claim 59 is positioned as the central node, surrounded by interconnected nodes representing the SOI components: larger teal nodes indicate annotated evidence directly related to the claim, smaller red nodes represent thematically related claims, and smaller teal nodes denote associated evidence linked to these related claims. Importantly, each component in the SOI is selectively included if relevant to Claim 59. For instance, while Evidence_59_2 and Evidence_59_3 are included, the remaining annotated evidence (from the total of six pieces for Claim 59 in the dataset) are excluded. Similarly, for the related Claim_1, only Evidence_1_2 and Evidence_1_3 are included, while the rest of its six associated evidence pieces are excluded. This selectivity highlights how this method prioritizes the most pertinent evidence and connections for Claim 59. This visualization underscores the rich thematic interconnections that the SOI provides, enhancing contextual understanding and facilitating more targeted evidence retrieval for the claim under investigation, as discussed in the subsequent text.

Following this preprocessing step of SOI identification, we introduced one of the core contributions of this work: constructing a thematic embedding for Claim 59 from the SOI, which serves as a key component of the query for evidence retrieval in the proposed CARAG framework. Specifically, we selected three key components from the SOI: annotated evidence, related claims, and thematic cluster evidence. Each of these components was then encoded using *all-mpnet-base-v2*. The individual embeddings were then aggregated through averaging, as outlined in Equation 7.1,

to create a unified thematic embedding via CEA that encapsulates the wider context of Claim 59 while intentionally excluding the claim itself.

This thematic embedding supports CARAG's context-aware approach by integrating both local and global perspectives, ensuring the influence of direct and contextual insights from the underlying corpus to inform evidence retrieval. This foundation not only enhances subsequent claim verification and post hoc explanations beyond instance-level local explainability but also advances the capabilities of traditional RAG methods.

Context-Aware Evidence Retrieval for Claim 59

Using the thematic embedding generated for CARAG, we conducted evidence retrieval for Claim 59, incorporating it as part of the retrieval query. To enable a comparative evaluation, we implemented three different retrieval approaches: (1) retrieving only the annotated evidence from *FactVer* as the ground truth evidence identified during dataset annotation; (2) applying the baseline RAG approach, which utilizes only the claim vector for evidence retrieval from the FAISS vectorized corpus (setting $\alpha = 1$ in Equation 7.2, as detailed in Section 7.4.2); and (3) using CARAG with a balanced combination of the claim vector and thematic embedding by setting $\alpha = 0.5$ in Equation 7.2.

For each approach, we selected the top $k = 6$ evidence items, in alignment with our dataset distribution statistics (Section 7.3.2), which indicate that the majority of claims are supported by six pieces of evidence. Table 7.3 presents a side-by-side comparison of evidence retrieved by these three approaches for Claim 59.

Table 7.3 Comparison of Annotated Evidence, Retrieved Evidence through RAG and CARAG, for Claim 59.

Annotated Evidence	RAG Retrieved Evidence	CARAG Retrieved Evidence
1. *Government pledge to act on the climate emergency.* 2. *U.N. notes 1.5 degrees Celsius as a crucial limit.* 3. *Todd's opposition to increasing traffic during a climate emergency.* 4. *Declaration of a climate emergency in 2020.* 5. *Groundswell NZ's protest against He Waka Eke Noa.* 6. *Chinese protests demanding action on climate change.*	1. *Failure will result in the country's once-successful car-making industry being scrapped.* 2. *Two-week journey which would have meant missing Wales' first match.* 3. *Shares fell as low as $6.50-apiece on Monday, down 97 percent from August 2021.* 4. *BMW-branded cars, motorcycles, and Mini models sold since October 1 get the new warranty.* 5. *Korean officials discussing several possible options to correct what they believe to be unfair policies that eliminated up to $7,500 of tax credits for EVs produced outside North America.* 6. *White House statement about Biden's health condition.*	1. *Failure will result in the country's once-successful car-making industry being scrapped.* 2. *Greenhouse gas trading scheme forms part of the UK government's ambition to achieve net-zero emissions by 2050.* 3. *Short-notice public investigatory attention affecting businesses.* 4. *Korean officials discussing options to correct unfair EV tax policies.* 5. *Two-week journey potentially missing Wales' first match.* 6. *Reliance on imports from China, US, and Europe for the car industry.*

A key observation from the evidence comparison in Table 7.3 is the overlap between certain evidence items retrieved by RAG and CARAG (e.g., references to the car-making industry and Korean EV tax policies). This overlap underscores CARAG's effectiveness in capturing a broad context similar to RAG while offering enhanced thematic alignment to the claim's topic. CARAG further strengthens this retrieval by incorporating additional climate-specific evidence directly related to the selected claim, demonstrating its advantage in filtering relevant information from broader contextual data.

Smart Prompting for Explanation Generation for Claim 59

Finally, we independently incorporated the evidence retrieved by each approach, into the LLM prompt to conduct the comparative analysis of explanation generation. This informed prompting (Section 7.4.3) supports evidence-based fact verification and explanation(post-hoc) generation, leveraging the previously introduced *Llama-2-7b-chat-hf* model.
The LLM prompt for each approach (annotated evidence, RAG, and CARAG) for Claim 59 is formatted as follows:

`Prompt: <Claim 59 (claim text)> + <K docs> + <specific instruction`[3]`>`

`<K docs>` is the only variable here, which corresponds to the retrieved evidence of each approach (representing the six retrieved evidence items ($k = 6$) selected for each approach). Specifically, for the annotated evidence approach, `<K docs>` refers to the items in the 'Annotated Evidence' column of Table 7.3; and for RAG and CARAG, `<K docs>` refers to the items in the ' RAG Retrieved Evidence' column and 'CARAG Retrieved Evidence' column of Table 7.3 respectively.

Figure 7.7 presents the generated explanations for each approach, aligned with the three types of prompts. For comprehensiveness, the figure also includes the claim text and its abstractive explanation, providing full context for the claim under investigation. Observations and limitations for each approach are highlighted, offering a thorough view of their respective strengths and constraints. Notably, all three explanations refute the claim, indicating it is not supported by the evidence.

The qualitative comparison in Figure 7.7 further classifies the explanations into local and global reasoning. Explanations based on annotated evidence (left) provide

[3]For example: You are a fact verification assistant. From the given Claim and its Evidence, determine if the claim is supported by the evidence and generate a concise explanation (two sentences max).

a direct assessment without broader context and are thus categorized as local reasoning. In contrast, the RAG and CARAG explanations, which incorporate a broader set of evidence to provide thematic perspectives beyond the immediate claim, fall under global reasoning. This distinction implies that, despite agreement in claim veracity, each approach offers a unique level of thematic depth. For instance, the RAG-generated explanation addresses broader economic aspects but lacks a direct thematic connection to the climate emergency, resulting in a more surface-level narrative.

By comparison, CARAG integrates climate-specific details with broader economic and governmental insights, offering a more comprehensive reflection of public and policy perspectives on climate issues. CARAG's approach leverages this global perspective effectively, balancing claim-specific elements with thematic coherence to enhance relevance and interpretability. This layered approach, connecting climate change to economic impacts and policy actions, demonstrates CARAG's ability to generate trustworthy explanations for nuanced, high-stakes claims by integrating broader, non-local context. This deeper contextual alignment surpasses RAG's capabilities, producing user-aligned explanations that encompass both thematic and factual nuances.

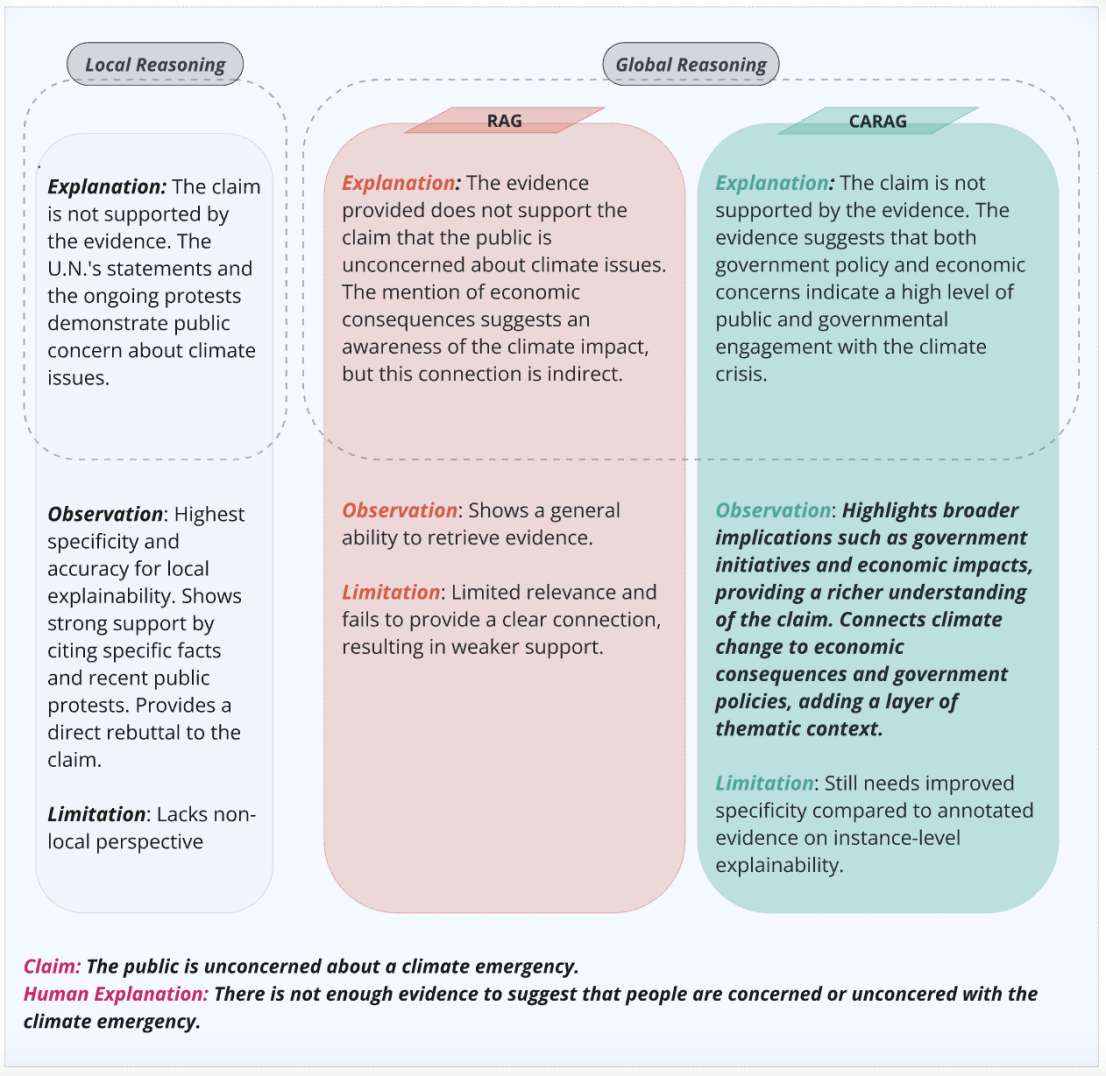


Fig. 7.7 Comparison of generated explanations across retrieval approaches for Claim 59, showcasing the methodology's application.

Through this case study, we underscore the dual benefits of CARAG: its proficiency in selecting contextually relevant evidence that deepens understanding and its capacity to translate this evidence into explanations that resonate with user expectations for interpretability and reliability. This analysis exemplifies how CARAG achieves balanced explainability by combining both local (claim-specific) and global (thematic) insights to provide a comprehensive and trustworthy explanation.

Moreover, CARAG leverages both textual and visual explanations, two widely recognized forms of XAI representation (Al-Ansari et al., 2024). As illustrated in Figure 7.6, Panel (b), visual explanations use graphical elements to clarify decision-making processes, while Figure 7.7 highlights CARAG's textual explanations, which

offer natural language reasoning that provides intuitive insights into the model's rationale. By aligning with these two forms of XAI, CARAG enhances both interpretability and transparency in fact verification, resulting in a comprehensive and insightful explainability mechanism.

In summary, CARAG's approach demonstrates superiority over RAG by providing a multi-faceted view that resonates with both the thematic and factual elements of the claim. To further substantiate these findings, in-depth comparative evaluation results of global explainability, focusing on RAG and CARAG across multiple claims, are presented in the upcoming section.

7.5.2 Comparative Analysis of RAG and CARAG Approaches

To evaluate CARAG's effectiveness in contrast to RAG, we focused on three critical aspects: contextual alignment, thematic relevance, and coverage, as key indicators of both local and global coherence. For this purpose, we conducted a comparative analysis across the three themes (COVID, Climate, and Electric Vehicles) in *FactVer*. For each theme, we generated post-hoc explanations for 10 claims using annotated evidence and both the RAG and CARAG approaches with adjustments to α in Equation 7.2, as demonstrated in the case study. This resulted in a total of 30 explanations per approach, organized in a CSV file for structured analysis, totaling 90 explanations across all themes. Our approach assesses the thematic alignment, coherence, and robustness of CARAG-generated explanations, using metrics such as density contours generated through kernel density estimation (KDE) for each theme and alignment comparison to that of RAG. These metrics are visualized through scatter plots and density contours to reveal the thematic depth and distribution of explanations produced by both RAG and CARAG.

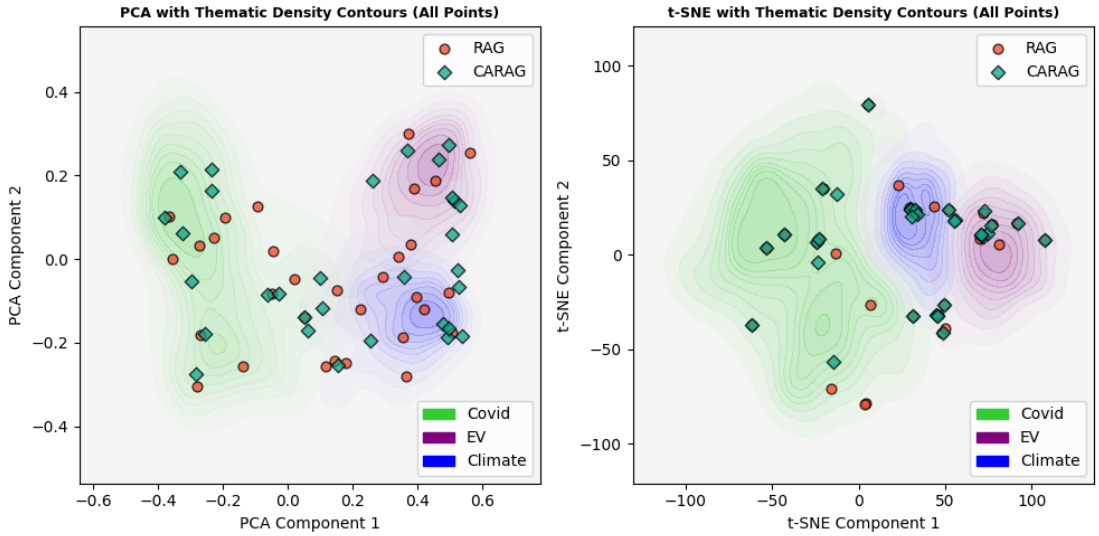
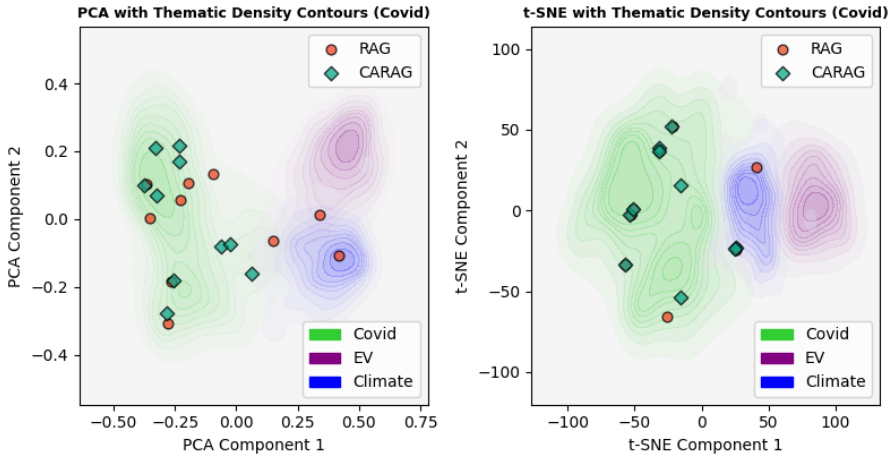


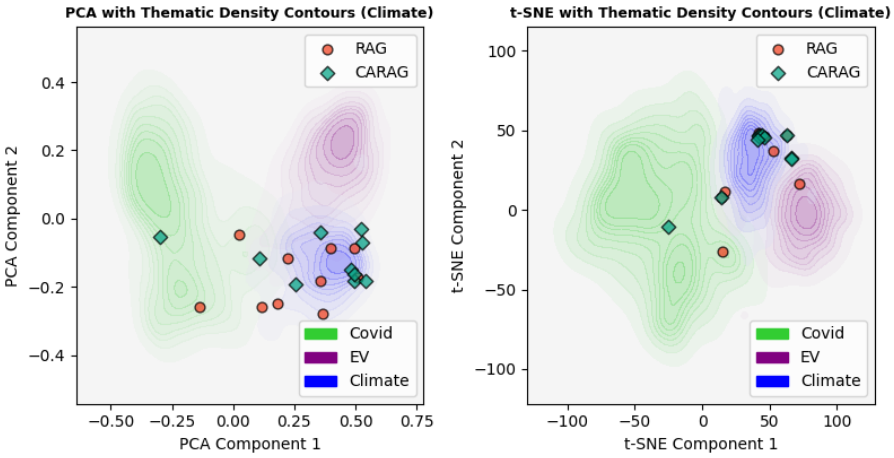
Fig. 7.8 PCA (left) and t-SNE (right) visualizations of embedding distributions for RAG-generated explanations (red circles) and CARAG-generated explanations (green diamonds), shown with KDE-based thematic density contours in the background (green for COVID, blue for Climate, and purple for Electric Vehicles). These contours illustrate thematic boundaries, enabling a comparative evaluation.

To facilitate an intuitive comparison of thematic clustering, we projected the embeddings of generated explanations into a 2D space using both PCA and t-SNE. The KDE-based density contours provide smooth, continuous representations of the thematic regions for each topic. Figure 7.8 presents an overview of all 30 explanations, with each point representing a RAG (red circles) or CARAG (green diamonds) generated explanation, plotted over density contours that illustrate thematic boundaries. These contours are color-coded by theme: green for COVID, blue for Climate, and purple for Electric Vehicles. This visualization provides a holistic view of how explanations from RAG and CARAG distribute across thematic contexts, with PCA (left) and t-SNE (right) visualizations.

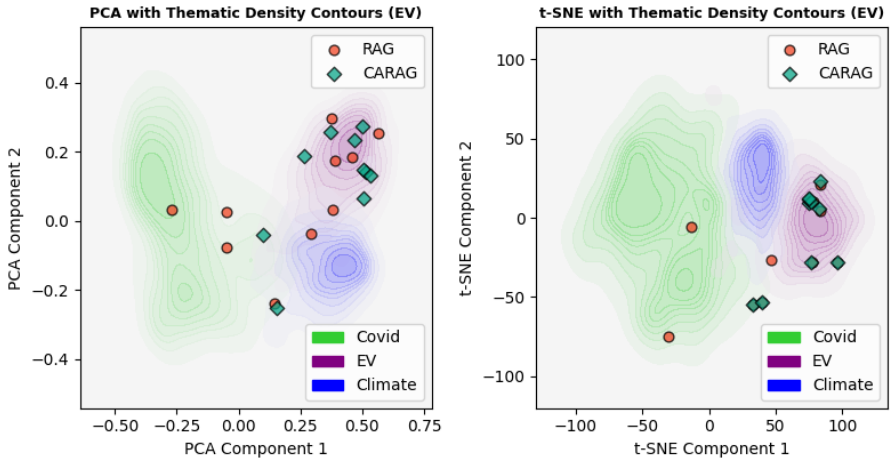
PCA reduces high-dimensional data to 2D while retaining the maximum variance, allowing us to observe broad distribution patterns, clusters, and outliers. This projection shows that RAG captures a generalized, global view, evident in its broader spread, but may lack theme-specific focus. t-SNE, conversely, better highlights local relationships and reveals tighter clusters around thematic boundaries, enhancing the interpretability of context-specific alignment. This view reveals that CARAG's explanations are more centrally aligned within each thematic area, suggesting a stronger focus on theme-specific context, while RAG explanations appear more peripheral, reflecting a broader, less targeted alignment.



(a) Embedding Distribution of Generated Explanations for COVID Theme



(b) Embedding Distribution of Generated Explanations for Climate Theme



(c) Embedding Distribution of Generated Explanations for Electric Vehicles Theme

Fig. 7.9 PCA (left) and t-SNE (right) visualizations of embedding distributions, highlighting how each approach aligns with underlying thematic regions.

To provide more granular insights into each theme, we present separate plots for each themes in Figure 7.9, showing the 10 explanation examples generated for each category, with contours for all themes included in each plot. This approach allows us to more clearly observe CARAG's ability to generate explanations that align with their corresponding thematic contours in the KDE representation. For example, in the COVID theme plot (Panel (a) in Figure 7.9), CARAG explanations cluster tightly within the green contour, indicating strong thematic alignment. Similarly, in the Climate (Panel (b)) and Electric Vehicles (Panel (c)) plots, CARAG explanations are concentrated within the blue and purple contours, respectively, underscoring CARAG's capacity for contextually relevant retrieval. While some RAG points do align within their respective theme contours, the majority are positioned along the periphery, suggesting a more generalized retrieval approach rather than theme-specific targeting. This difference highlights CARAG's superior ability to produce explanations with closer thematic alignment, enhancing context-specific relevance.

RAG's distribution reveals a tendency to capture generalized information across themes, which aligns with its retrieval-augmented nature but may dilute thematic specificity. Conversely, CARAG's thematic retrieval is more focused, producing explanations that closely align with each theme's contours. By leveraging KDE-based density contours, CARAG explanations demonstrate tighter clustering within the intended thematic regions, underscoring its potential for theme-specific retrieval. This makes CARAG particularly suitable for tasks where contextual alignment is crucial, such as verifying claims in COVID-related topics, where thematic relevance enhances accuracy. The individual theme plots further illustrate this difference, showing that CARAG explanations are more concentrated within thematic contours, demonstrating enhanced thematic relevance compared to RAG.

Table 7.4 Quantitative comparison of RAG and CARAG embedding distributions across themes. Each sub-table shows distances to centroids in PCA and t-SNE spaces, with differences highlighted in the Diff (PCA) and Diff (t-SNE) columns. Per-theme averages are included as the last row in each sub-table.

(a) COVID Theme

Index	RAG (PCA)	CARAG (PCA)	Diff (PCA)	RAG (t-SNE)	CARAG (t-SNE)	Diff (t-SNE)
0	0.1133	0.2263	0.1130	31.5851	31.5608	-0.0242
1	0.1423	0.2506	0.1083	31.4120	31.4665	0.0544
2	0.3000	0.2720	-0.0280	31.4419	31.4444	0.0025
3	0.1913	0.1995	0.0083	31.4158	31.4030	-0.0128
4	0.0645	0.1354	0.0709	31.5440	31.4492	-0.0948
5	0.1779	0.1736	-0.0043	31.4712	31.4843	0.0132
6	0.1809	0.1767	-0.0042	31.6878	31.5558	-0.1320
7	0.6405	0.1687	-0.4717	32.1598	31.6895	-0.4704
8	0.3688	0.1994	-0.1693	31.9011	31.7242	-0.1769
9	0.5524	0.3178	-0.2346	32.0992	31.8005	-0.2987
Average	**0.2732**	**0.2120**	**-0.0612**	**31.6718**	**31.5578**	**-0.1140**

(b) Climate Theme

Index	RAG (PCA)	CARAG (PCA)	Diff (PCA)	RAG (t-SNE)	CARAG (t-SNE)	Diff (t-SNE)
0	0.0590	0.6626	0.6035	37.3316	37.8603	0.5287
1	0.2180	0.1234	-0.0946	37.5164	37.2024	-0.3139
2	0.0531	0.1442	0.0911	37.2505	37.2061	-0.0443
3	0.1518	0.2534	0.1016	37.3610	37.5232	0.1623
4	0.1551	0.1397	-0.0154	37.1891	37.1950	0.0060
5	0.1368	0.0855	-0.0512	37.4192	37.2654	-0.1538
6	0.2749	0.1878	-0.0871	37.5750	37.1636	-0.4115
7	0.3473	0.1930	-0.1543	37.5689	37.1082	-0.4607
8	0.1461	0.1789	0.0328	37.1552	37.1223	-0.0329
9	0.5143	0.1252	-0.3891	37.8044	37.4247	-0.3797
Average	**0.2056**	**0.2094**	**0.0037**	**37.4171**	**37.3071**	**-0.1100**

(c) Electric Vehicles Theme

Index	RAG (PCA)	CARAG (PCA)	Diff (PCA)	RAG (t-SNE)	CARAG (t-SNE)	Diff (t-SNE)
0	0.1989	0.1155	-0.0834	76.1226	76.1748	0.0522
1	0.6796	0.1405	-0.5391	76.9621	76.4237	-0.5385
2	0.4984	0.1093	-0.3891	76.7383	76.1809	-0.5574
3	0.1112	0.1338	0.0226	76.3114	76.1575	-0.1540
4	0.4627	0.4668	0.0041	76.5504	76.5409	-0.0095
5	0.0260	0.1643	0.1382	76.2979	76.1871	-0.1108
6	0.1565	0.1188	-0.0377	76.3120	76.3179	0.0059
7	0.2151	0.1403	-0.0748	76.3986	76.1808	-0.2178
8	0.4607	0.3522	-0.1085	76.7354	76.5901	-0.1453
9	0.0718	0.1159	0.0441	76.2318	76.2199	-0.0119
Average	**0.2881**	**0.1858**	**-0.1023**	**76.4661**	**76.2973**	**-0.1687**

The quantitative results in Table 7.4 corroborate the visual patterns observed in Figure 7.9, providing statistical evidence of CARAG's superior alignment with thematic regions. These results are based on Euclidean distances between the embeddings of RAG and CARAG explanations and the thematic centroids in PCA and t-SNE spaces. As shown in Table 7.4, for each theme, CARAG demonstrates consistently lower average distances to thematic centroids compared to RAG, particularly in t-SNE space, where the differences are more pronounced. Specifically, the differences (Diff(PCA) and Diff(t-SNE)) are calculated as
Diff (PCA or t-SNE) = CARAG Distance − RAG Distance. Negative values in the difference columns indicate CARAG's superior alignment (shorter distance to the center compared to RAG), highlighting its tighter clustering within thematic regions, and are color-coded in green. Positive values, color-coded in red, represent the rare instance where RAG outperformed CARAG, such as the Diff (PCA) for Climate. In contrast, likely due to its non-linear dimensionality reduction approach compared to PCA's linear reduction (an investigation into this aspect is planned for future work), t-SNE consistently highlights CARAG's tighter alignment. This numerical validation underscores CARAG's ability to maintain thematic specificity, with smaller distance variations highlighting its tighter clustering within the intended thematic regions. The inclusion of overall averages (calculated as averages of per-theme averages) in Table 7.5 provides a holistic view of CARAG's thematic alignment advantage, further demonstrating its ability to produce explanations that are more closely aligned with thematic contours compared to RAG.

Table 7.5 Overall averages of RAG and CARAG embedding distributions (last row), computed as averages of per-theme averages in PCA and t-SNE spaces, with color-coded highlights for performance.

Theme	RAG (PCA) Avg	CARAG (PCA) Avg	Diff (PCA) Avg	RAG (t-SNE) Avg	CARAG (t-SNE) Avg	Diff (t-SNE) Avg
Covid	0.2732	0.2120	**-0.0612**	31.6718	31.5578	**-0.1140**
Climate	0.2056	0.2094	**0.0037**	37.4171	37.3071	**-0.1100**
Electric Vehicles	0.2881	0.1858	**-0.1023**	76.4661	76.2973	**-0.1687**
Overall Average	**0.2556**	**0.2024**	**-0.0532**	**48.5183**	**48.3874**	**-0.1309**

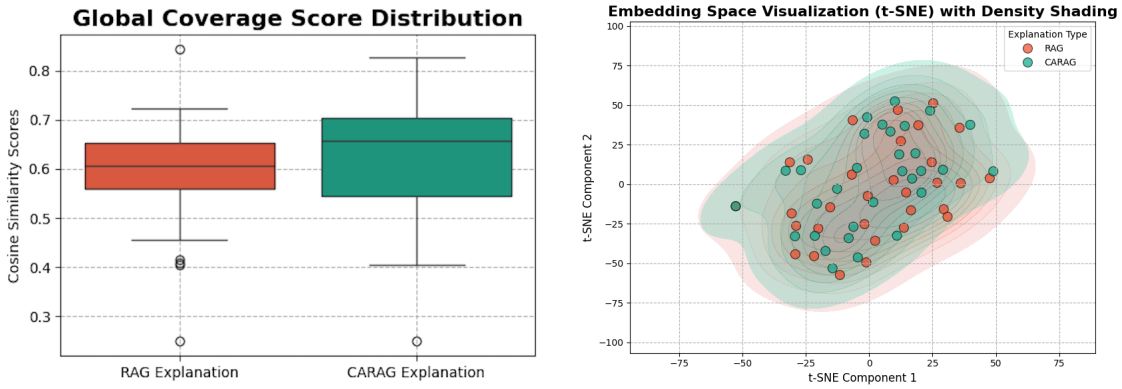
In summary, RAG offers broad-spectrum context suitable for general claims, while CARAG excels in generating thematically aligned, contextually precise explanations. This distinction highlights CARAG's potential for theme-specific fact

verification tasks, making it particularly effective in domains requiring context alignment, as demonstrated by its stronger alignment within each theme.

7.5.3 Limitations of Standard Analysis & Visualization Techniques in Explainable AI

Evaluating CARAG's integration of local and global perspectives in post-hoc explanations requires more than standard metrics and visualizations, which often fall short of capturing nuanced thematic and contextual relevance. Metrics like precision, recall, F1, MRR, and MAP measure retrieval performance but do not assess thematic alignment, a critical element in our framework. Similarly, overall accuracy and F1 scores capture binary prediction accuracy without addressing the thematic coherence of explanations. Moreover,standard explainability metrics, such as fidelity, interpretability scores, and sufficiency, typically offer insights at the individual explanation level, lacking the layered depth needed for complex thematic datasets. For instance, when examining the CARAG explanation in Figure 7.7, which emphasizes a rich thematic alignment by connecting climate change with economic impacts and government policies, it is clear that traditional metrics would not adequately capture this depth of thematic integration. Additionally, even similarity measures struggle here, as the CARAG-generated explanation provides context that aligns with thematic patterns beyond surface-level similarity, contrasting with the simpler human explanation in Figure 7.7, which lacks this layered thematic framing.

Standard visualization techniques, such as box plots, provide a limited view of CARAG's thematic alignment by reducing it to a numeric similarity measure. For example, Figure 7.10(a) shows a box plot of global coverage scores, where cosine similarity scores between CARAG's explanations and dataset vectors are calculated to gauge relevance. Although useful for assessing general alignment, this approach treats thematic coherence as a basic numeric metric, failing to capture the contextual depth CARAG aims to provide. Similarly, a t-SNE visualization with Kernel Density Estimation, as shown in Figure 7.10(b), highlights clustering within the embedding space without indicating clear thematic boundaries. Unlike our PCA and t-SNE approach in Section 7.5.2, which incorporates distinct KDE representations to define thematic contours, this generic t-SNE with KDE does not offer indicators of thematic relevance, making it insufficient for evaluating CARAG's context-aware framework.



(a) Global Coverage Score Distribution (b) t-SNE with Density Shading

Fig. 7.10 Standard visualization techniques: (a) Global coverage score distribution using cosine similarity, which does not provide insights into thematic relevance, and (b) t-SNE with KDE shading, lacking explicit thematic boundaries for RAG (red) and CARAG (green) explanations.

In summary, these standard techniques lack the nuanced depth necessary to evaluate CARAG's thematic alignment, highlighting the need for a tailored evaluation approach. For complex datasets where thematic contours are crucial, customized visualizations like our PCA and t-SNE contour-based method in Section 7.5.2, offer a more suitable, though still approximate, approach for capturing the multi-dimensional thematic relevance and contextual alignment central to CARAG's explainable AI goals. This underlines the importance of developing specialized evaluation measures for frameworks like CARAG, an area we aim to expand in future research.

7.6 Challenges and Limitations

In AFV, explainability can be approached from three primary perspectives, architectural, methodological, and data (Vallayil et al., 2023). While CARAG primarily contributes to the methodological (CARAG) and data (FactVer) aspects, these improvements fall short of addressing broader model-level interpretability, as noted in studies like Zhao et al. (2024a), which explore neuron activation patterns and component-specific functionalities. Approaches such as probing, neuron activation analysis, and mechanistic interpretation illuminate individual component functions, aiming to reveal latent structures within language models. CARAG, in contrast, emphasizes contextual alignment and transparency in retrieval logic and explanation

generation over internal model insights, distinguishing it from these model-centric interpretability methods.

Thus, while CARAG enhances explainability within the AFV pipeline, expanding to include model-level analysis remains an area for potential growth. Integrating such insights could offer a comprehensive understanding of both retrieval rationale and the latent knowledge embedded within the model itself, aligning CARAG more closely with a holistic view of explainability in AFV.

Another limitation of CARAG at this stage is its reliance on the FactVer dataset for evaluation. While CARAG's retrieval mechanism and prompting strategy are already data-agnostic, the SOI generation process still depends on thematic labels. To address this, we are working on enhancing the SOI generation and subsequent thematic embedding processes to function independently of pre-defined themes. These updates will enable CARAG to generalize across diverse datasets, with preliminary results to be presented in an upcoming publication.

7.7 Future Research Directions

Outlined below are the forthcoming steps aimed at further refining CARAG's retrieval and explainability capabilities:

1. **Label-Independent SOI Refinement**: We aim to eliminate reliance on theme labels for SOI refinement, enabling CARAG to generate SOIs without predefined themes.

2. **Comprehensive Ablation Study on Parameter Effects**: We plan to perform an in-depth ablation study to assess CARAG's component contributions by analyzing key parameters in SOI composition and retrieval vector generation. This analysis aims to determine how these parameters influence retrieval quality, thematic relevance, and interpretability.

3. **Agreement-Based Performance Evaluation**: Building on our analysis (Section 7.5.2), we aim to explore CARAG's alignment with human-annotated evidence using agreement metrics. Recognizing that standard evaluation methods may fall short in capturing thematic depth, we will experiment with tailored metrics to better assess CARAG's nuanced thematic alignment, ensuring transparency and reliability.

Extending beyond our immediate plans, CARAG's broader applications illustrate promising research directions for future exploration by the community. CARAG's capacity for thematic clustering and contextual visualization enables it to serve high-stakes fields such as investigative, legal, and policy analysis. By revealing non-local patterns of intent, misinformation dissemination, and behavioral inconsistencies across posts, CARAG supports a nuanced approach to fact verification. Furthermore, in longitudinal analyses, CARAG's macro-level perspective can be instrumental in identifying evolving misinformation trends, empowering agencies and policymakers with insights critical for developing long-term strategies to enhance public awareness and media literacy.

7.8 Conclusion

CARAG stands out as an explainable AI framework by integrating evidence-based claim verification with post-hoc explanation generation in a transparent and interpretable manner. Unlike traditional fact verification approaches, which focus narrowly on annotated evidence and often yield highly localized insights, or the highly global RAG, which retrieves evidence without explicitly revealing the rationale behind each retrieval choice, CARAG structures its retrieval query by combining the claim vector with the SOI vector. This enables context-aware evidence retrieval grounded in clear, interpretable logic.

CARAG's transparency is further enhanced by its visual interpretability: the SOI graph provides a map of the components influencing the retrieval process and illustrates a network of thematically interconnected information centered around the claim, offering clear visual insight into the SOI components that enhance retrieval transparency. Notably, the absence of some annotated evidence within this network at times underscores the specificity and intentionality of our approach, distinguishing CARAG from conventional strategies.

Additionally, by handling retrieval and generation as distinct steps rather than as a single-stage process (as in standard RAG), CARAG offers deeper insight into why specific evidence is selected and how it contributes to optimized prompting for the generation pipeline. This modular approach's flexibility is achieved through two hyperparameters that influence distinct stages: a threshold parameter, δ, for refining the SOI based on similarity, and an adjustable parameter, α, for balancing the influence of the claim vector against thematic embeddings in the final vector combination.

Furthermore, CARAG is supported by FactVer, a novel, explanation-focused dataset specifically curated to enhance thematic alignment and transparency in AFV. FactVer provides both local and global perspectives by pairing claims with multiple annotated evidence entries in various thematic contexts, advancing research in explainability-focused AFV studies and laying a strong foundation for CARAG's nuanced, context-aware approach.

Together, these elements make CARAG a promising advancement toward a more interpretable and contextually aware framework, bringing distinct layers of explainability to the AFV pipeline (as illustrated in Figure 7.5). By enhancing both the methodological and data perspectives of XAI in AFV, CARAG and FactVer collectively reinforce transparency and reliability, addressing gaps left by traditional methods and setting a robust path for more explainable AFV systems.

Chapter 8

Prelude - Manuscript 4

Building on the context-aware retrieval introduced in Manuscript 3, which established CARAG, this section explores the next step: extending thematic discovery to unsupervised settings. While CARAG integrates predefined thematic clusters and SOIs to enhance local and global explainability in AFV, its reliance on annotated themes limits adaptability to datasets lacking structured labels.

Manuscript 4 (Chapter 9) introduces CARAG-u, an unsupervised extension of CARAG, designed to overcome this limitation. Instead of relying on predefined thematic labels, CARAG-u dynamically identifies semantic clusters, allowing claims and evidence to self-organize based on emergent relationships, autonomously forming SOIs without requiring external annotations.

This advancement enables CARAG-u to generalize beyond structured datasets, making explainable AFV more adaptable to open-domain verification. By preserving CARAG's transparency while eliminating its dependence on labeled themes, CARAG-u represents the first step toward a scalable, data-agnostic approach to thematic retrieval in AFV.

In practice, applying CARAG-u to open-domain corpora requires substantial but tractable computational resources. Experiments in this thesis were conducted on an NVIDIA GeForce RTX 4080 with 16GB GDDR6X VRAM, which supported embedding extraction, clustering, and LLaMA-7B inference. The 7B model operates within the capacity of a single 16GB GPU when quantized, but larger models (>7B) demand progressively greater resources, with full-precision deployments already exceeding the capabilities of commodity GPUs. This underscores an important trade-off between model scale, efficiency, and accessibility, suggesting that practical open-domain deployments may depend on quantization strategies, multi-GPU setups, or future hardware advances.

Chapter 9

Unsupervised Thematic Context Discovery for Explainable AI in Fact Verification: Advancing the CARAG Framework (Manuscript 4)

9.1 Introduction

Post-hoc explanations (Moradi and Samwald, 2021) have become a widely adopted solution in Explainable Artificial Intelligence (XAI), aiming to clarify the decisions of complex deep learning models, yet ironically, they often rely on equally complex models like Large Language Models (LLMs) for generating these explanations. This reliance underscores the trade-off between leveraging state-of-the-art generative capabilities and ensuring interpretability, particularly in Automated Fact Verification (AFV), where trust and transparency in evidence-based reasoning are critical. Alongside LLMs, Retrieval Augmented Generation (RAG) frameworks (Lewis et al., 2020) have gained traction for their ability to dynamically retrieve relevant evidence for fact verification, making them highly adaptable across various fact-checking scenarios. RAG retrieves facts from an external knowledge base to feed LLMs during the generative process. This creates a multi-layered challenge for XAI in AFV: while these sophisticated systems excel in advanced retrieval and generative capabilities, they inherently lack transparency, particularly in how evidence is selected and how this influences the generated explanation, underscoring the need for innovative methodologies to ensure interpretability and reliability.

Addressing this challenge, the Context-Aware Retrieval Augmented Generation (CARAG) framework (Vallayil et al., 2025) was introduced as a step toward enhancing explainability in AFV. It provides an approach to interpreting both the evidence retrieval process and the post-hoc explanations generated using the retrieved evidence. CARAG achieves this by enhancing the evidence retrieval query; instead of relying primarily on claim (query) embeddings, as is conventional in many RAG systems, it incorporates thematic context alongside the claim embedding to enrich the retrieval process. This modification significantly influences evidence selection and has been empirically proven to improve the thematic alignment of the claim with the generated post-hoc explanations. By doing so, CARAG interprets and enriches evidence selection, thereby enhancing post-hoc explanations and contributing to advancements in addressing critical XAI challenges.

However, CARAG derives its thematic embeddings from a predefined subset of the fact verification dataset, which is dynamically determined through statistical modeling and semantic aggregation. While this approach enhances transparency in evidence selection and the relevance of generated explanations, it is inherently constrained by its dependence on theme/topic annotations and claim-evidence pair labels. These structured annotations serve as the foundation for CARAG's thematic embedding generation, limiting its applicability to datasets that are already annotated. This reliance not only restricts CARAG's utility in open-domain or unstructured datasets but also highlights its limitations in scaling to broader, annotation-free scenarios.

In this paper, we introduce CARAG-u, an enhanced framework that eliminates reliance on structured annotations by dynamically deriving thematic clusters and evidence pools in an unsupervised manner. This advancement broadens CARAG-u's applicability to unstructured datasets, enabling seamless operation without predefined labels, extending its usability to open-domain settings. To evaluate its effectiveness, we benchmark CARAG-u against RAG, while acknowledging CARAG as an enhancement of RAG. Despite operating independently of pre-annotated labels, CARAG-u surpasses the RAG baseline and demonstrates competitive performance with CARAG as shown in Tables (9.1a) & (9.1b). Crucially, CARAG-u achieves this advancement while preserving CARAG's core explainability features, thereby addressing a key challenge in scaling XAI solutions for AFV systems.

For evaluation, we use the same FactVer_v2.0 dataset as employed in CARAG, available on HuggingFace[1]. Although FactVer includes theme and claim-evidence

[1] https://huggingface.co/datasets/manjuvallayil/factver_master

annotations, these annotations are not utilized during evidence retrieval or explanation generation in CARAG-u. Instead, they are used solely to evaluate performance, particularly to assess the thematic alignment of the generated explanations, ensuring a consistent baseline and fair comparison with CARAG. This approach isolates the impact of transitioning from a supervised to an unsupervised framework while leveraging a dataset with known properties to assess CARAG-u's scalability, thematic discovery capabilities, and relevance in evidence-based reasoning tasks. Building on these design considerations, the CARAG-u framework is designed with scalability, ensuring adaptability to advancements in XAI, RAG, and LLMs. Its modular architecture allows seamless integration of state-of-the-art techniques from LLM research and RAG innovations with minimal adaptation, keeping the framework at the forefront of explainability in AFV. The complete CARAG-u implementation, is publicly available on GitHub[2].

It is equally important to highlight that both CARAG and CARAG-u enhance transparency in AFV by addressing the critical challenge of integrating local and global XAI concepts. In the context of AFV, local explainability focuses on clarifying individual predictions, whereas global explainability encompasses diverse approaches to understanding the model's overall reasoning behavior, thereby offering a more holistic view of its decision-making logic. By integrating these perspectives, CARAG and CARAG-u provide deeper insights into how individual claims relate to the broader context of a knowledge base, where context plays a pivotal role in interpreting individual facts. However, prominent literature reviews and surveys in the intersection of XAI and AFV (Kotonya and Toni, 2020a; Vallayil et al., 2023) highlight persistent gaps in this field, especially the limited focus on achieving global transparency. Existing XAI approaches in AFV, such as transformer-based summarization (Atanasova et al., 2020; Kotonya and Toni, 2020b), logic-based models (Chen et al., 2022a; Krishna et al., 2022), attention mechanisms (Amjad et al., 2023; Popat et al., 2018; Shu et al., 2019), counterfactual explanations (Dai et al., 2022; Xu et al., 2023), and methods leveraging RAG for dynamic evidence retrieval and reasoning (Singhal et al., 2024; Wang and Shu, 2023), predominantly focus on local explainability, leaving the broader challenges of achieving global transparency in AFV systems largely unaddressed. Recent surveys on LLM-based fact checking (Vykopal et al., 2024) highlight the potential of LLMs to support fact-checkers with advanced reasoning capabilities, but they do not directly address the challenge of achieving global transparency in AFV systems. To the best of our knowledge, the

[2]https://github.com/manjuvallayil/factver_dev

existing literature highlights the CARAG framework (Vallayil et al., 2025), along with its precursor work on graph-based thematic clustering for explainability in AFV (Vallayil et al., 2024), as the only prior efforts in related work explicitly addressing the integration of global explainability in AFV. These works uniquely combine a claim's local context with its position within the dataset's global context. While not directly focused on global explainability, methodological parallels in the literature can be drawn to broader XAI research, such as surrogate models like LIME (Ribeiro et al., 2016), which approximate complex AI model behaviors locally using Machine Learning (ML) models to provide human-understandable instance-level explanations. In a similar vein, CARAG leverages interpretable ML-based methods to illuminate the decisions of complex AI models, bridging the gap between advanced model performance and the need for interpretability in AFV systems.

The remainder of this paper is organized as follows. The methodology section details the CARAG-u methodology, including dynamic thematic context discovery, query embedding construction, and a side-by-side depiction of CARAG and CARAG-u in Figure 9.1. The Experiments and Results section presents the experimental setup, with comparative evaluation against RAG and CARAG, and highlights CARAG-u's evidence-based reasoning performance through a case study. Finally, the Discussion section summarizes our findings and outlines directions for future work.

9.2 Methodology

This section details the steps involved in dynamically identifying thematic clusters and generating retrieval query embeddings, enabling evidence retrieval and explanation generation without prior annotations.

9.2.1 Dynamic Thematic Context Discovery

The methodology begins by representing the dataset $\mathcal{D}$, encompassing claims and evidences, in a unified semantic space using sentence embeddings. These embeddings capture semantic relationships across dataset elements, providing a foundation for subsequent clustering to discover thematic contexts.

Clustering is performed using a Gaussian Mixture Model (GMM) optimized via the Expectation-Maximization (EM) algorithm, as shown in Equation 9.1. GMM-EM was chosen for its capacity to model the dataset as a mixture of latent thematic patterns, where each pattern is represented as a Gaussian component characterized

by its mean and variance (Al-Dujaili Al-Khazraji and Ebrahimi-Moghadam, 2024; Barai et al., 2022; Jiao et al., 2023; Moondra and Chahal, 2023). This probabilistic formulation enables soft clustering, which is suitable for capturing overlapping and ambiguous themes in natural language. It dynamically identifies a set of t clusters ($\{C_i\}_{i=1}^{t}$) based on inherent semantic relationships. The parameter t, representing the number of clusters, is configurable depending on the dataset and desired thematic resolution.

$$L = \text{GMM-EM}(\{\text{emb}(e_i)\}_{i=1}^{n},\ t) \rightarrow \{C_i\}_{i=1}^{t} \tag{9.1}$$

where L represents the cluster labels assigned to each embedding, and $\text{emb}(e_i)$ denotes the embedding of the i-th textual input from dataset $\mathcal{D}$. This step eliminates the reliance on predefined thematic filtering $\mathcal{T}$, used in CARAG to create $\mathcal{D}_T$ prior to clustering, marking an initial progression towards an unsupervised approach, as illustrated in the *Clustering Phase* of Figure 9.1. To ensure consistency and enable a fair comparison with CARAG during evaluation, the same SBERT model (sentence-transformers/all-mpnet-base-v2) and clustering algorithm (sklearn.mixture.GaussianMixture) were selected, leveraging their established effectiveness in semantic representation (Reimers and Gurevych, 2019) and clustering (Binti Kasim et al., 2021), respectively.

Subsequently, the cluster C_k containing the embedding of the selected claim c_s is identified from these clusters. A Subset of Interest (SOI), is then constructed from C_k by selecting evidences e_k that meet a similarity threshold δ, as defined in Equation 9.2, alongside c_s itself.

$$\mathcal{S}(C_k) = \{c_s\} \cup \{e_k \mid \text{sim}(e_k, c_s) > \delta \text{ and } e_k \in C_k\} \tag{9.2}$$

where, $\mathcal{S}(C_k)$ represents the SOI for c_s, comprising the claim c_s and thematically relevant evidences e_k from C_k.

This approach differs fundamentally from CARAG, which relies on annotated evidences explicitly tied to c_s, as well as related claims c_r within the cluster and their corresponding annotated evidences referred to as “thematic_cluster_evidence”, as illustrated in the *Soi generation Phase* of Figure 9.1. In contrast, CARAG-u considers all evidences within the cluster C_k as the evidence pool, irrespective of their claim annotations, initially referred to as “cluster_evidence". This pool serves as the basis for SOI formation, which is further refined into “refined_cluster_evidence” by applying the similarity threshold δ. By eliminating the dependency on claim-

evidence annotations, CARAG-u enables an unsupervised and scalable process. This dynamic SOI formation allows CARAG-u to generalize the CARAG framework, operating effectively on datasets without thematic labels or predefined structures. By leveraging unsupervised processes, CARAG-u enhances its applicability and adaptability for thematic discovery.

9.2.2 Evidence Retrieval Query Construction from Discovered Contexts

Building on the thematic context identified through $\mathcal{S}(C_k)$, this step integrates the discovered context into the retrieval query for fetching relevant evidence from the fact-checking dataset. To achieve this, a thematic embedding T_e is first computed as the average embedding of the refined cluster evidences within $\mathcal{S}(C_k)$, encapsulating the cluster-level context for c_s.

CARAG-u then proceeds to construct a combined embedding E_{combined}, integrating thematic insights from T_e with claim-specific focus from E_{claim} (also denoted interchangeably as $emb(c_s)$) to form the retrieval query, as defined in Equation 9.3.

$$\begin{aligned} E_{combined} &= \alpha \cdot E_{claim} + (1-\alpha) \cdot T_e, \\ T_e &= \frac{1}{m}\sum_{j=1}^{m} emb(e_j), \\ e_j &\in \mathcal{S}(C_k)[\text{'}refined\_cluster\_evidences\text{'}] \end{aligned} \tag{9.3}$$

The weighting parameter α (a user-defined parameter) controls the balance between claim-specific precision and thematic context, offering flexibility for diverse retrieval scenarios. While the mathematical formulation of E_{combined} aligns with CARAG, CARAG-u differs by deriving $\mathcal{S}(C_k)$ and T_e from a dynamically formed, unsupervised cluster-level evidence pool, as discussed in the previous sub section.

The resulting E_{combined}, which serves as the evidence retrieval query, ensures that the retrieved evidences from $\mathcal{D}$ are aligned with both the claim c_s and the thematic context discovered.

9.2.3 Pipeline Summary and Implementation

Algorithm 1 outlines the methodology, detailing the steps for clustering, SOI formation, thematic embedding generation, and evidence retrieval using the combined embedding as the query. The practical application of this algorithm is demonstrated

Algorithm 1 Evidence Retrieval with CARAG-u

Require: Dataset $\mathcal{D}$, Selected Claim c_s, Similarity Threshold δ, Number of Evidence Docs to retrieve n_{docs}, Weighting Factor α, Number of Clusters t

Ensure: Top n_{docs} evidences retrieved from $\mathcal{D}$ for the selected claim.

Step 1: Data Preparation and Clustering

Load dataset $\mathcal{D}$

Initialize empty list *all_texts* and append all claims and evidences in $\mathcal{D}$

for each element $e \in$ *all_texts* **do**

    Generate embedding $\text{emb}(e)$

end for

Apply GMM-EM to cluster $\text{emb}(\textit{all\_texts})$:
$L = \text{GMM-EM}(\{\text{emb}(e_i)\}_{i=1}^{n},\ t) \rightarrow \{C_i\}_{i=1}^{t}$

Assign cluster labels $L = \{l_1, l_2, \ldots, l_n\}$ to all texts in *all_texts*, where $l_i \in \{c_i, c_j, \ldots, c_t\}$ and t is the number of clusters.

Step 2: SOI Generation for Selected Claim

Determine the cluster C_k containing c_s

Initialize SOI dictionary

Extract all evidences in C_k to temporary list 'cluster_evidences' (evidence pool)

for each evidence e_k in cluster_evidences **do**

    Compute similarity:

$$\text{sim} = \text{cosine\_similarity}(E_{claim}, \text{emb}(e_k)),$$
$$E_{claim} = \text{emb}(c_s)$$

    if $\text{sim} > \delta$ **then**

        Add e_k to the list for key 'refined_cluster_evidences' in SOI

    end if

end for

Step 3: Thematic Embedding Generation

Compute thematic embedding:

$$T_e = \frac{1}{m} \sum_{j=1}^{m} \text{emb}(e_j),$$
$$e_j \in \text{SOI['refined\_cluster\_evidences']}$$

Step 4: Evidence Retrieval

Compute the Combined Embedding:

$$E_{\text{combined}} = \alpha \cdot E_{claim} + (1 - \alpha) \cdot T_e$$

Retrieve top n_{docs} evidences using E_{combined}:

$$\text{retrieved\_evidence} = \text{Retriever}(E_{\text{combined}}, n_{\text{docs}})$$

Output: The top-n_{docs} evidences for c_s retrieved from $\mathcal{D}$ using CARAG-u.

in the case study detailed in the Experiments and results section, showcasing its adaptability to varying parameter configurations. Additionally, the case study highlights the post-hoc explanations generated from the retrieved evidences, underscoring CARAG-u's effectiveness in unsupervised thematic discovery.

Figure 9.1 summarizes the distinctions between CARAG and CARAG-u across the clustering, SOI generation, and thematic embedding phases. While CARAG operates within predefined thematic subsets, CARAG-u bypasses theme-based filtering and directly applies clustering to the entire dataset ($\mathcal{D}$), enabling dynamic cluster formation without relying on thematic annotations. In the SOI generation phase, CARAG incorporates annotated evidences, related claims, and thematic cluster evidences, whereas CARAG-u solely relies on refined cluster evidences derived from the unsupervised clustering process. An experimental evaluation of our methodology is presented in the following section.

9.3 Experiments & Results

We adopt a RAG-based pipeline wherein initial evidence retrieval is performed using FAISS (Douze et al., 2024) on sentence embeddings generated by the *all-mpnet-base-v2* variant of Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). This ensures semantically meaningful and scalable retrieval from our document corpus. The explanation generation is powered by the *Llama-2-7b-chat-hf* model (Touvron et al., 2023a), implemented via Hugging Face's Transformers library using a sequence-to-sequence prompting template that combines the claim, retrieved evidence, and an instructional query. All experiments are conducted in zero-shot mode. To ensure consistency, the same embedding and generation settings used in the original CARAG framework are retained. The following subsections present CARAG-u's performance in generating post-hoc explanations through both qualitative and quantitative evaluations. First, we conduct a visual analysis of the explanation embeddings followed by a quantitative assessment using thematic alignment metrics. Finally, we include a focused case study that illustrates how dynamic thematic embeddings influence evidence retrieval and the resulting explanation quality.

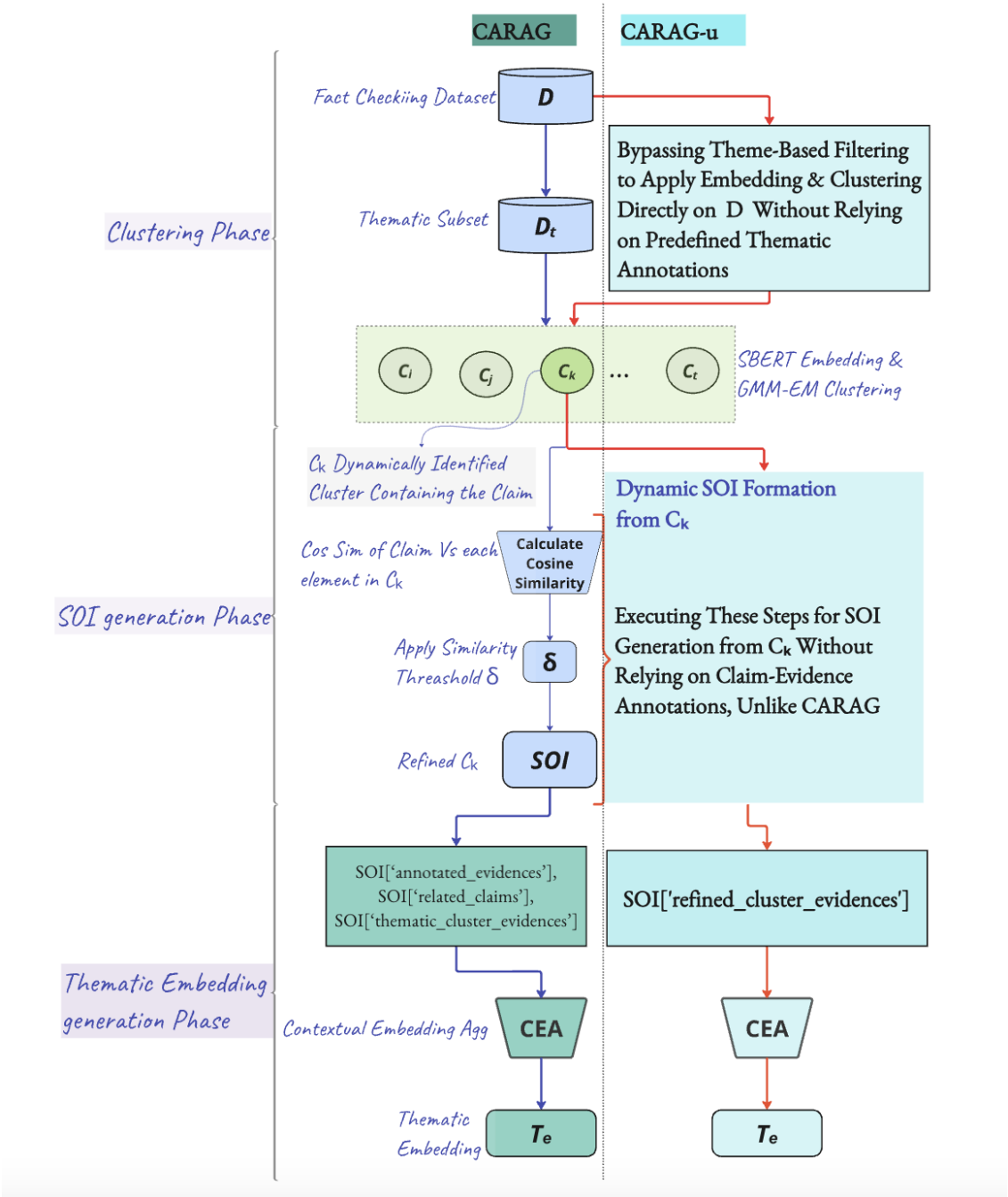


Fig. 9.1 Comparison of CARAG and CARAG-u across clustering, SOI generation, and thematic embedding phases. CARAG relies on predefined thematic subsets and annotated evidence, while CARAG-u dynamically clusters the entire dataset ($\mathcal{D}$), forming SOIs and thematic embeddings without relying on either thematic annotations or claim-evidence annotations.

9.3.1 Evaluating Thematic Alignment Across RAG, CARAG, and CARAG-u

Building on the methodological framework detailed, we evaluate the effectiveness of CARAG-u in generating contextually relevant explanations. For both CARAG and CARAG-u, we first generated thematic embeddings T_e based on the respective SOIs. Using these T_e, we constructed combined embeddings E_{combined} as retrieval queries and subsequently retrieved n_{docs} evidences for a selected claim. Post-hoc explanations were then generated using these evidence sets, formatted as part of the LLM prompt[3], with the *Llama-2-7b-chat-hf* (Touvron et al., 2023a) operating in a zero-shot paradigm.

As part of this evaluation, we selected 10 claims each from three themes, COVID, Climate, and Electric Vehicles. For each claim, evidence documents were first dynamically retrieved from the FactVer dataset using RAG, and then using E_{combined}, computed with T_e corresponding to the CARAG and CARAG-u approaches. This process ultimately resulted in a total of 90 post-hoc explanations. FactVer was chosen for its unique structure, specifically designed to address gaps in existing AFV datasets (Kotonya and Toni, 2020a) by supporting both verification and explainability research with structured evidence relationships and multi-evidence claims. Unlike datasets such as FEVER (Thorne et al., 2018a) or MultiFC (Augenstein et al., 2019), which focus primarily on fact verification, FactVer's multi-evidence structure supports XAI research initiatives in the AFV domain, particularly in advancing post-hoc explanation generation approaches. This makes it suitable for evaluating frameworks like CARAG-u, while also fostering broader progress in XAI for AFV systems.

Having generated post-hoc explanations for comparative evaluation, we now discuss the baselines used to benchmark CARAG-u's performance. While CARAG provided the foundational framework for CARAG-u, RAG serves as the baseline to ensure a fair comparison. RAG employs a generalized retrieval strategy, operating solely on the global evidence pool without thematic filtering or clustering. In contrast, CARAG benefits from supervised thematic filtering, clustering, and annotated evidence to generate thematic embeddings. Consequently, using CARAG as a baseline would not provide an unbiased assessment of CARAG-u's unsupervised capabilities. Instead, comparing CARAG-u to RAG highlights its adaptability and thematic robustness in the absence of predefined annotations.

[3] prompt: <claim> + <n_{docs} evidence documents> + <instruction specifying the evidence-based claim verification and post-hoc explanation generation tasks>

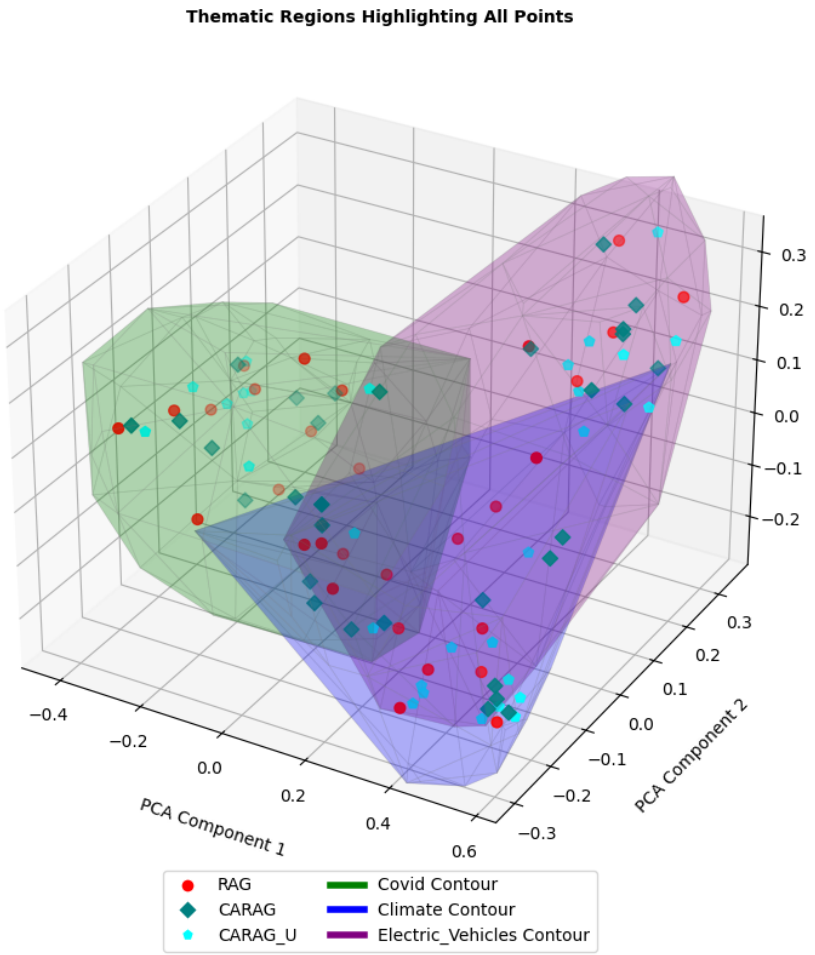
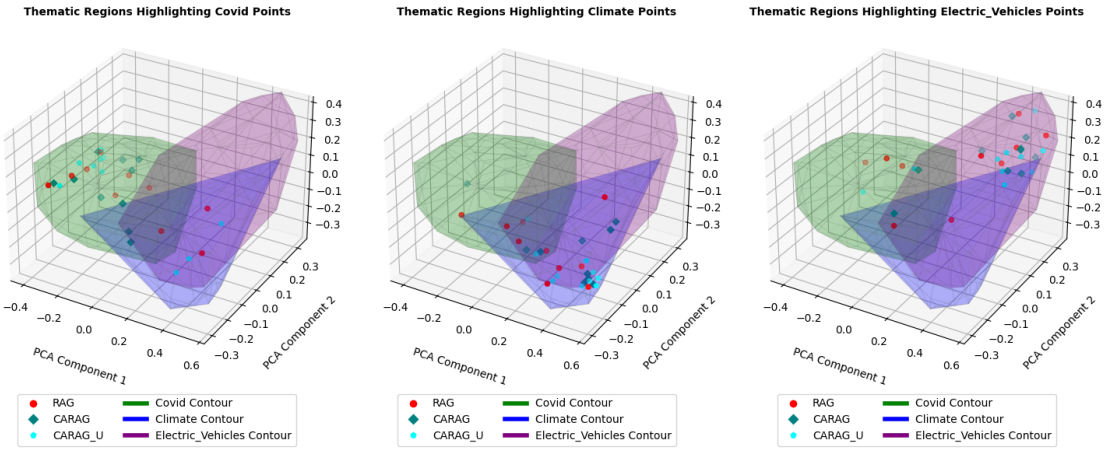


Fig. 9.2 Visualization of explanation embeddings for RAG, CARAG, and CARAG-u within thematic boundaries (COVID: green, Climate: blue, Electric Vehicles: purple).

Figure 9.2 provides an overall visualization of the embeddings of the post-hoc explanations generated across the three frameworks. Thematic boundaries (COVID: green, Climate: blue, Electric Vehicles: purple) are depicted using 3D PCA-based convex hulls, delineating the thematic regions. RAG explanations (red circles) are broadly scattered, reflecting the generalized nature of its retrieval strategy. CARAG explanations (teal diamonds) exhibit tighter clustering within thematic boundaries due to its reliance on supervised thematic annotations. Notably, CARAG-u explanations (cyan pentagons) demonstrate comparable alignment within thematic regions, despite operating in an unsupervised manner. This demonstrates CARAG-u's ability to dynamically infer thematic context, highlighting its flexibility in generating contextually relevant explanations without annotated evidence.



(a) COVID (b) Climate (c) Electric Vehicles

Fig. 9.3 Explanation embeddings disaggregated by theme (COVID, Climate, Electric Vehicles). Each subplot illustrates the alignment of RAG, CARAG, and CARAG-u explanations within thematic boundaries.

Figure 9.3 further disaggregates the analysis by theme, with each subplot showing explanation embeddings for one theme plotted over all thematic boundaries. For example, Figure 9.3a shows that the explanations generated for COVID claims predominantly fall within the green contour. CARAG explanations remain tightly clustered, reflecting their reliance on thematic supervision. Similar patterns are observed for Climate (Figure 9.3b) and Electric Vehicles (Figure 9.3c), further underscoring CARAG-u's ability to approximate the performance of CARAG while operating without evidence and thematic annotations.

Table 9.1 Alignment of explanation embeddings with thematic centroids across PCA and t-SNE spaces.

(a) Euclidean distances between explanation embeddings (RAG, CARAG, and CARAG-u) and their respective thematic centroids across PCA and t-SNE spaces for each theme, including average values. These distances measure how closely the embeddings align with their thematic regions, with lower values indicating better alignment.

Theme	#	RAG (PCA)	CARAG (PCA)	CARAG-u (PCA)	RAG (t-SNE)	CARAG (t-SNE)	CARAG-u (t-SNE)
Climate	1	0.0642	0.6879	0.1626	14.9536	15.6389	15.0701
	2	0.2651	0.1297	0.1502	15.1499	14.8257	14.8050
	3	0.0703	0.1532	0.1613	14.9061	14.8084	14.7986
	4	0.1567	0.2484	0.1562	14.9309	15.1907	14.8102
	5	0.1712	0.1475	0.1824	14.7886	14.8085	14.7711
	6	0.1349	0.0958	0.0830	15.0864	14.9705	14.9422
	7	0.3511	0.1957	0.2076	15.2261	14.7633	14.7499
	8	0.3488	0.2748	0.1191	15.2973	14.8389	14.9329
	9	0.5537	0.1451	0.2875	15.4724	15.0653	15.2112
	10	0.3884	0.2557	0.1742	14.9002	14.8288	14.8327
	Avg	0.2505	0.2334	0.1684	15.0712	14.9739	14.8924
COVID	1	0.1218	0.2302	0.2409	12.7828	12.7467	12.5792
	2	0.1414	0.2630	0.2941	12.6108	12.6549	12.5617
	3	0.3442	0.3105	0.3051	12.6562	12.6574	12.6926
	4	0.1923	0.2008	0.2190	12.6017	12.5889	12.5870
	5	0.1529	0.2331	0.1700	12.7568	12.6671	12.6365
	6	0.2194	0.2087	0.1658	12.6747	12.6908	12.5963
	7	0.2119	0.1896	0.1601	12.8909	12.7509	12.6086
	8	0.6613	0.1771	0.6123	13.3939	12.9051	13.3338
	9	0.3976	0.2770	0.6175	13.1286	12.9587	13.3131
	10	0.5557	0.3579	0.6253	13.3042	13.0402	13.3754
	Avg	0.2998	0.2448	0.3410	12.8801	12.7661	12.8284
EV	1	0.2399	0.1741	0.1352	33.9902	34.0469	34.0512
	2	0.6843	0.1378	0.1425	34.8393	34.2922	34.0821
	3	0.5027	0.1596	0.1255	34.6234	34.0523	34.1068
	4	0.1630	0.1375	0.2075	34.1918	34.0296	33.9812
	5	0.4811	0.4711	0.6347	34.4458	34.4383	34.7523
	6	0.0227	0.1657	0.1567	34.1666	34.0517	34.0275
	7	0.2334	0.2198	0.2671	34.1765	34.1851	34.1106
	8	0.2948	0.1588	0.1564	34.2807	34.0566	34.0586
	9	0.4585	0.3520	0.3715	34.6112	34.4718	34.4926
	10	0.1092	0.1582	0.0461	34.1030	34.0882	34.1630
	Avg	0.3190	0.2135	0.2243	34.3429	34.1713	34.1826

(b) Differences in average distances (CARAG minus RAG and CARAG-u minus RAG) across PCA and t-SNE spaces for each theme. Negative values indicate better alignment compared to the baseline (RAG), with larger negative values representing greater improvement.

Theme	CARAG-RAG (PCA)	CARAG-u-RAG (PCA)	CARAG-RAG (t-SNE)	CARAG-u-RAG (t-SNE)
Climate	-0.0171	-0.0820	-0.0973	-0.1788
COVID	-0.0551	0.0412	-0.1140	-0.0516
EV	-0.1055	-0.0946	-0.1716	-0.1603
Overall Average	-0.0592	-0.0452	-0.1276	-0.1302

Tables (9.1a) and (9.1b) complement these visualizations with quantitative insights. Table 9.1a reports Euclidean distances between explanation embeddings (RAG, CARAG, and CARAG-u) and their respective thematic centroids across PCA and t-SNE spaces for each theme, including average values. These distances measure how closely the embeddings align with their thematic regions, with lower values indicating better alignment. CARAG-u consistently achieves tighter alignment than RAG across most themes and embedding spaces. Notable improvements are observed for Climate, where CARAG-u achieves the lowest t-SNE distance (14.8924), while demonstrating competitive performance with CARAG for COVID and Electric Vehicles. The inclusion of t-SNE distances highlights CARAG-u's ability to capture local relationships, enhancing its contextual alignment. Table (9.1b) shows differences in alignment (CARAG minus RAG & CARAG-u minus RAG distance in PCA and in t-SNE spaces). Negative values indicate superior alignment over RAG. CARAG-u exhibits substantial improvements in Climate (-0.082 in PCA, -0.1788 in t-SNE) and Electric Vehicles (-0.0946 in PCA, -0.1603 in t-SNE). While CARAG benefits from supervised thematic embeddings, CARAG-u performs competitively, as evidenced by overall averages (-0.0452 for PCA and -0.1302 for t-SNE). In summary, CARAG-u balances unsupervised adaptability with thematic alignment, offering an alternative to RAG in contexts where structured annotations are unavailable. These results validate its robustness and extend the explainability framework established by CARAG to broader, less structured domains, advancing the goal of unsupervised fact verification.

9.3.2 Case Study

Building on the broader framework comparison, we conducted a focused case study to assess CARAG-u's performance in dynamically generating thematic embeddings and their influence on evidence retrieval and explanation generation. This evaluation, centered on varying the weighting parameter α (Equation 9.3), used the claim, "Climate change is caused by atmospheric concentration of greenhouse gases" (Claim ID: 44)[4] from the FactVer dataset.

The case study followed the methodological steps outlined in Algorithm 1, with the following parameter configuration: $\mathcal{D} = FactVer$, $c_s = 44$, $\delta = 0.75$, $n_{\text{docs}} = 6$, $\alpha \in \{1, 0.75, 0.50, 0.25\}$, and $t = 10$. Anchoring the evaluation to this structured process highlights CARAG-u's adaptability to varying parameter configurations,

[4]Claim selected for its global relevance and prominence as a highly discussed topic.

Table 9.2 Case study example of post-hoc explanations generated by CARAG-u for the claim "Climate change is caused by atmospheric concentration of greenhouse gases." These explanations are based on retrieved evidence for different α values (Equation 9.3), with *delta* set to 0.75 (Equation 9.2), and t set to 10 (Equation 9.1). The variations in α demonstrate its impact on thematic alignment and evidence retrieval. The retrieved evidence sets corresponding to each configuration are provided in Appendix B.1.

Configuration	Post-hoc Explanation
$\alpha = 1, \delta = 0.75, t = 10$	**Explanation:** "The claim is not supported by the evidence. While the evidence mentions the impact of long COVID on the labor market, it does not provide any information on the concentration of greenhouse gases in the atmosphere and their role in causing climate change."
Observation: In this configuration ($\alpha = 1$), the framework behaves as RAG, retrieving evidence solely based on the claim text (E_{claim}). This results in a thematically inconsistent evidence set, often containing unrelated details.	
$\alpha = 0.75, \delta = 0.75, t = 10$	**Explanation:** "The claim is not supported by the evidence provided. While the evidence suggests that long COVID can affect the labor market, it does not provide enough data to support the claim that 1.6 million full-time equivalent workers could be out of work due to long COVID."
Observation: Evidence retrieval begins to incorporate thematic elements, though precision remains inconsistent, blending some relevant and irrelevant information.	
$\alpha = 0.5, \delta = 0.75, t = 10$	**Explanation:** "The claim is supported by the evidence provided. The report from Brookings Metro estimates that 1.6 million full-time equivalent workers could be out of work due to long COVID, which supports the claim that the condition could have a significant impact on the labor market."
Observation: Thematic embedding significantly influences retrieval, but the precision of claim specificity is reduced.	
$\alpha = 0.25, \delta = 0.75, t = 10$	**Explanation:** "The evidence provided supports the claim that climate change is caused by the atmospheric concentration of greenhouse gases. The UK's industrial decarbonization schemes and emissions trading highlight the role of greenhouse gases in climate change."
Observation: Evidence retrieval heavily favors thematic embedding (T_e), providing evidence highly aligned with the claim. Precision in claim specificity improves.	

where α determines the extent to which the thematic context T_e is incorporated into the evidence retrieval query E_{combined} as defined in Equation 9.3. Consequently, when $\alpha = 1$, CARAG-u relies solely on E_{claim} for evidence retrieval, mimicking the behavior of a traditional RAG framework as our baseline setup. Conversely as α decreases, the influence of T_e increases, incorporating broader thematic context (global context) at the expense of claim specificity.

Table 9.2 illustrates the evolution of post-hoc explanations generated by CARAG-u across varying α configurations. At α = 1 or at the baseline RAG setup, producing explanations disconnected from the thematic context of the claim (climate change), such as an irrelevant focus on labor market impacts. As α decreases, the thematic embedding T_e increasingly influences retrieval process, aligning the retrieved evidence and consequently the explanations, more closely with the claim. Notably, at α = 0.25, the explanations emphasize industrial de-carbonization schemes and greenhouse gas emissions, directly supporting the claim. While its predecessor CARAG also supports adaptability in evidence retrieval through varying α, CARAG-u stands out by achieving this in a fully unsupervised manner.

9.4 Discussion

CARAG-u advances XAI for AFV by integrating thematic context into evidence retrieval and explanation generation, enhancing transparency and relevance through dynamically computed SOIs, without relying on structured annotations or predefined thematic labels. Despite these advancements, challenges remain. Specifically, its adaptability to datasets with highly heterogeneous thematic structures requires further investigation to assess its performance and limitations in such contexts. Additionally, ensuring its post-hoc explanations are interpretable to non-expert users remains a broader challenge for XAI systems. Addressing these limitations will strengthen CARAG-u's applicability. This direction is further reinforced by a recent survey (Vykopal et al., 2024), which aimed to advance the understanding and integration of LLMs in AFV. The survey highlights that knowledge-augmented strategies such as RAG remain significantly underutilized in fact-checking, with only a small fraction of surveyed works incorporating external sources. This identified gap reinforces the relevance of our approach, which explores how RAG-based systems can also advance transparency in AFV.

Our future work on CARAG-u will focus on expanding its adaptability and interpretability through key enhancements. Adaptive clustering techniques will

be explored to dynamically determine the optimal number of clusters (t), enabling improved scalability across datasets with varying thematic structures. Additional evaluations on datasets beyond FactVer will assess CARAG-u's robustness across diverse domains, while a systematic analysis of hyperparameters, including δ, n_{docs}, and t, aims to refine evidence retrieval and enhance the contextual relevance of explanations. Future evaluations on datasets with diverse structures will further validate CARAG-u's scalability and its capacity to adapt to heterogeneous thematic complexities, addressing broader research objectives in explainable AFV.

Chapter 10

Conclusion and Future Directions

This thesis contributes to XAI in AFV by addressing critical limitations in explainability, introducing novel methodologies, and extending retrieval-augmented verification frameworks. Through the **CARAG** and **CARAG-u** frameworks, alongside the **FactVer** dataset, this research advances both **theoretical understanding** and **practical implementation** of explainable AFV.

10.1 Key Contributions

1. **Systematic Evaluation of Explainability in AFV (Manuscript 1):** This thesis began with a comprehensive review of explainability challenges in AFV, identifying gaps in local and global transparency and dataset limitations (RQ1).

2. **Thematic Discovery and Context Visualization (Manuscript 2):** To address the lack of systematic methods for identifying and representing global thematic relationships, this work introduced the SOI methodology, enabling a structured view of claim-related thematic clusters (RQ2).

3. **Development of CARAG and FactVer (Manuscript 3):** Building on SOI-based thematic discovery, CARAG was introduced to enhance retrieval-augmented verification by integrating thematic embeddings into evidence retrieval (RQ3). FactVer, an explanation-focused fact-checking dataset, was also developed to support explainability studies in AFV.

4. **Unsupervised Explainability with CARAG-u (Manuscript 4):** To address CARAG's reliance on thematic labels and evidence annotations, CARAG-u was introduced as an unsupervised extension that dynamically identifies

thematic clusters, marking an initial step towards adaptability in open-domain verification (RQ4).

10.2 Impact and Future Directions

This thesis advances explainability in AFV by integrating context-aware thematic discovery, retrieval-augmented verification, and unsupervised explainability. These contributions enhance transparency by bridging local and global perspectives, offering a more interpretable fact verification pipeline. FactVer, introduced as a benchmark dataset, supports structured evaluation of thematic alignment and explanation-focused methodologies and is publicly available on Hugging Face for broader research use. Meanwhile, CARAG and CARAG-u provide adaptable frameworks for both structured and open-domain verification, with implementations made accessible through GitHub. Beyond academic contributions, this research has implications for high-stakes applications such as misinformation detection, policy analysis, and legal AI, where trust and interpretability are essential.

While this work establishes a strong foundation, several avenues remain for future exploration. Adaptive clustering techniques could refine thematic discovery, optimizing SOI formation for datasets with varying thematic structures. Further research into human-centered evaluation metrics could improve interpretability by aligning AI-generated explanations with user needs, while explanation viability may also be assessed using established measures such as BLEU, ROUGE, BERTScore, and Perplexity, as well as BETA metrics including Fidelity, Unambiguity, and Interpretability Size. Finally, enhancing scalability and real-time adaptability would allow CARAG-u to function effectively in dynamic verification environments. As a step toward practical adoption, the methodological contributions of this research are being prepared for release as a Python package, facilitating seamless integration into explainable fact verification pipelines and fostering further development in the field.

Moreover, this work satisfies the principles of Responsible and Ethical AI by enhancing transparency through interpretable evidence selection, reducing bias via unsupervised thematic discovery that lessens reliance on human-labeled data, and fostering accountability by making model reasoning traceable in high-stakes fact verification contexts. By integrating context-aware retrieval and unsupervised thematic discovery, this research lays the groundwork for more interpretable, scalable, and adaptable AFV systems, contributing to the broader goals of trustworthy AI.

10.3 Final Reflections

This research represents a step forward in making AFV systems both explainable and adaptable, ensuring that AI-driven fact verification is transparent, scalable, and context-aware. By bridging foundational insights with methodological innovations, this work lays the groundwork for future advancements in XAI, addressing the growing need for trust and accountability in AI-driven decision-making.

Beyond the technical advancements, this research reinforces the necessity of explainability-focused datasets and methodologies, recognizing that AFV transparency extends beyond algorithmic improvements to structured evaluation frameworks. The introduction of FactVer, CARAG, and CARAG-u provides not only theoretical contributions but also practical tools for researchers and practitioners working toward interpretable and reliable fact verification.

More broadly, this thesis contributes to the ongoing discourse on trustworthy AI, emphasizing the balance between retrieval performance, interpretability, and adaptability in verification pipelines. While challenges remain, such as improving real-time adaptability, refining evaluation metrics, and scaling thematic discovery across open-domain fact verification, this research lays a foundation for future advancements. As AI-driven fact verification continues to evolve, explainability must remain a core priority, ensuring that verification systems are not only accurate but also transparent, interpretable, and aligned with real-world needs. This thesis aspires to contribute toward that vision, providing an adaptable framework that bridges algorithmic sophistication with explainability-centered design.

References

Ahmadi, N., Lee, J., Papotti, P., and Saeed, M. (2019). Explainable Fact Checking with Probabilistic Answer Set Programming. In *Conference for Truth and Trust Online*.

Al-Ansari, N., Al-Thani, D., and Al-Mansoori, R. S. (2024). User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Human Behavior and Emerging Technologies*, 2024(1):4628855.

Al-Dujaili Al-Khazraji, M. J. and Ebrahimi-Moghadam, A. (2024). An Innovative Method for Speech Signal Emotion Recognition Based on Spectral Features Using GMM and HMM Techniques. *Wireless Personal Communications*, 134(2):735–753.

Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805.

Amjad, H., Ashraf, M. S., Sherazi, S. Z. A., et al. (2023). Attention-Based Explainability Approaches in Healthcare Natural Language Processing. In *Proceedings of the International Conference on Health Informatics (HEALTHINF)*, pages 689–696.

AnthropicAI (2023). Introducing Claude. Retrieved from https://www.anthropic.com/index/introducing-claude.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Barai, B., Chakraborty, T., Das, N., Basu, S., and Nasipuri, M. (2022). Closed-Set Speaker Identification Using VQ and GMM Based Models. *International Journal of Speech Technology*, 25(1):173–196.

Binti Kasim, F. A., Pheng, H. S., Binti Nordin, S. Z., and Haur, O. K. (2021). Gaussian Mixture Model - Expectation Maximization Algorithm for Brain Images. In *2021*

2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), pages 1–5.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*.

Brailas, A., Koskinas, K., Dafermos, M., and Alexias, G. (2015). Wikipedia in Education: Acculturation and learning in virtual communities. *Learning, Culture and Social Interaction*, 7:59–70.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Chen, J., Bao, Q., Sun, C., et al. (2022a). Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.

Chen, J., Zhang, R., Guo, J., Fan, Y., and Cheng, X. (2022b). GERE: Generative Evidence Retrieval for Fact Verification. In *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2184–2189. Association for Computing Machinery, Inc.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Inc., Thousand Oaks, CA, 3rd edition.

Dai, S. C., Hsu, Y. L., Xiong, A., and Ku, L. W. (2022). Ask to Know More: Generating Counterfactual Explanations for Fake Claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2800–2810.

Das, A., Liu, H., Kovatchev, V., and Lease, M. (2023). The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

DeHaven, M. and Scott, S. (2023). BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2024). The FAISS Library. *CoRR*, abs/2401.08281.

Du, Y., Bosselut, A., and Manning, C. D. (2022). Synthetic Disinformation Attacks on Automated Fact Verification Systems. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gad-Elrab, M. H., Stepanova, D., Urbani, J., and Weikum, G. (2019). Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 87–95.

Gardner, M. and Mitchell, T. (2015). Efficient and expressive knowledge base completion using subgraph feature extraction. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498.

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42.

Gunning, D. (2016). Broad agency announcement explainable artificial intelligence (XAI). Technical report, Technical report, Defense Advanced Research Projects Agency Information Innovation Office 675 North Randolph Street Arlington, VA 22203-2114.

Gunning, D., Vorm, E., Wang, J. Y., and Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Hanselowski, A., Stab, C., Schulz, C., Li, Z., and Gurevych, I. (2019). A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., and Tremayne, M. (2017). ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proc. VLDB Endow.*, 10(12):1945–1948.

Huang, Y., Gao, M., Wang, J., and Shu, K. (2021). DAFD: Domain Adaptation Framework for Fake News Detection. In Mantoro, T., Lee, M., Ayu, M. A., Wong, K. W., and Hidayanto, A. N., editors, *Neural Information Processing*, pages 305–316, Cham. Springer International Publishing.

Iliadis, D., Peikou, M., Adamidou, C., and Kyriakopoulou, A. (2024). A comparison of embedding aggregation strategies in drug–target interaction prediction. *BMC Bioinformatics*, 25(1):59.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jayanthi, S. M., Embar, V., and Raghunathan, K. (2021). Evaluating Pretrained Transformer Models for Entity Linking in Task-Oriented Dialog. In Bandyopadhyay, S., Devi, S. L., and Bhattacharyya, P., editors, *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 537–543, Silchar, India. NLP Association of India (NLPAI), National Institute of Technology Silchar.

Jiang, K., Pradeep, R., and Lin, J. (2021). Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, volume 2, pages 402–410.

Jiao, Z., Ji, Y., Gao, P., and Wang, S. H. (2023). Extraction and Analysis of Brain Functional Statuses for Early Mild Cognitive Impairment Using Variational Auto-Encoder. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12.

Kim, T. W. (2018). Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test. *ArXiv*, abs/1810.09598.

Kotonya, N. and Toni, F. (2020a). Explainable Automated Fact-Checking: A Survey. In *28th International Conference on Computational Linguistics, Proceedings of the Conference (COLING)*, pages 5430–5443. Online.

Kotonya, N. and Toni, F. (2020b). Explainable Automated Fact-Checking for Public Health Claims. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Krishna, A., Riedel, S., and Vlachos, A. (2022). ProoFVer: Natural Logic Theorem Proving for Fact Verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020a). On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Li, X., Zhong, J., Wu, X., Yu, J., Liu, X., and Meng, H. (2020b). Adversarial Attacks on GMM i-Vector Based Speaker Verification Systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6579–6583. IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liusie, A., Manakul, P., and Gales, M. J. F. (2024). LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. *arXiv preprint arXiv:2307.07889*. To Appear at EACL 2024.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Moondra, A. and Chahal, P. (2023). Speaker Recognition Improvement for Degraded Human Voice Using Modified-MFCC with GMM. *International Journal of Advanced Computer Science and Applications*, 14(6).

Moradi, M. and Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165:113941.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. Technical report, DARPA XAI Program, 40 S. Alcaniz St., Pensacola, FL 32502.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Olivares, D. G., Quijano, L., and Liberatore, F. (2023). Enhancing Information Retrieval in Fact Extraction and Verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 38–48.

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods*. SAGE Publications, Newbury Park, CA, 2nd edition.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Popat, K., Mukherjee, S., Yates, A., and Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 22–32.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., and Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Rani, A., Tonmoy, S. M. T. I., Dalal, D., Gautam, S., Chakraborty, M., Chadha, A., Sheth, A., and Das, A. (2023). FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, pages 1135–1144.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215.

Salemi, A. and Zamani, H. (2024). Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, pages 2395–2400. Association for Computing Machinery.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2698–2706.

Shi, B. and Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133.

Shorten, C., Khoshgoftaar, T. M., and Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of Big Data*, 8(1):1–54.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Singhal, R., Patwa, P., Patwa, P., Chadha, A., and Das, A. (2024). Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Soleimani, A., Monz, C., and Worring, M. (2019). BERT for evidence retrieval and claim verification. In *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science*, volume 12036 LNCS, pages 359–366.

Stammbach, D. (2021). Evidence Selection as a Token-Level Prediction Task. In *FEVER 2021 - Fact Extraction and VERification, Proceedings of the 4th Workshop*, pages 14–20. Association for Computational Linguistics (ACL).

Stammbach, D. and Ash, E. (2020). e-FEVER: Explanations and Summaries for Automated Fact Checking. *BRISK Binary Robust Invariant Scalable Keypoints*, pages 12–19.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR, pages 3319–3328.

Tang, J., Yin, D., Zhang, J., and Yin, J. (2023). Secure Embedding Aggregation for Federated Representation Learning. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2392–2397. IEEE.

Thorne, J. and Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018b). The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Touvron, H., Martin, L., Stone, K., et al. (2023a). LLAMa-2: Open foundation and fine-tuned chat models. Retrieved from https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Touvron, H., Martin, L., Stone, K. R., et al. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.

Vallayil, M., Nand, P., and Yan, W. Q. (2024). Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems. In *ACIS 2024 Proceedings*, number 101 in ACIS Proceedings Series.

Vallayil, M., Nand, P., Yan, W. Q., and Allende-Cid, H. (2023). Explainability of Automated Fact Verification Systems: A Comprehensive Review. *Applied Sciences*, 13(23):12608.

Vallayil, M., Nand, P., Yan, W. Q., Allende-Cid, H., and Vamathevan, T. (2025). CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI. *Applied Sciences*, 15(4):1970.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009. Neural information processing systems foundation.

Vykopal, I., Pikuliak, M., Ostermann, S., and Simko, M. (2024). Generative Large Language Models in Automated Fact-Checking: A Survey. *ArXiv*, abs/2407.02351.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wang, H. and Shu, K. (2023). Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wiegreffe, S. and Marasovic, A. (2021). Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Xu, W., Liu, Q., Wu, S., and Wang, L. (2023). Counterfactual Debiasing for Fact Verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024a). Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Zhao, W., Zheng, W., Wang, Y., and Wang, T. (2024b). Fake News Detection Based on Knowledge-Guided Semantic Analysis. *Electronics*, 13(2):259.

Zhong, W., Xu, J., Tang, D., Xu, Z., Duan, N., Zhou, M., Wang, J., and Yin, J. (2020). Reasoning over semantic-level graph for fact checking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180. Association for Computational Linguistics (ACL).

Appendix A

A.1 Template and Instructions for Annotation

Table A.1 Fields Used by Annotators for Claim and Evidence Generation

Field	Description
Claim ID	An integer ID allocated to each claim.
Evidence ID	A unique ID for each evidence, ranging from E1 to E6, to ensure traceability between claims and their corresponding evidence.
Claim	A span of text specifying the statement of fact or assertion to be verified.
Label	A label indicating the veracity of the claim—True (T), False (F), or Not Enough Info (N).
Evidence	The text span that supports, refutes, or remains neutral about the claim.
Article ID	The identifier from the ***FactVer_1.1*** dataset containing the evidence text span.
Reason	A text description providing the rationale for the assigned label, which could be novel or derived from one of the evidence spans.

A.1.1 Steps for Annotation

1. **Obtain the Template:** Obtain the Excel template file containing fields as specified in Table A.1.

2. **Select Topic:** Choose the allocated topic and filter the FactVer_1.1 dataset to only show articles on your chosen topic.

3. **Generate Claims:** Read an article from the filtered set and generate a claim in the Excel file. Copy and paste sentences that you think support the claim as evidence. Label the claim-evidence pair as either True (T), False (F), or Not Enough Info (N).

4. **Record Article ID:** Copy the article ID from the FactVer_1.1 dataset into your spreadsheet in the "Article ID" field.

5. **Gather Additional Evidence:** Read other or the same articles for further evidence to either support, refute, or be neutral about your claim.

6. **Create Evidence Rows:** Make 6 rows for each claim, each containing a different evidence span. If no additional evidence is found, fill up with neutral text to reach 6 rows per claim. Typically, a claim is expected to have multiple supporting evidence pieces and some neutral evidence. It is also possible for a claim to have some supporting and some refuting evidence.

A.1.2 Notes on Claim and Evidence Creation

1. **Evidence-Based Labeling:** Claims should be labeled according to the evidence in the dataset, without considering outside knowledge.

2. **Consistent Labeling:** Each claim must have a consistent label (T, F, or N) across all 6 rows. While the evidence set may contain text spans that contradict the claim, the label should reflect the overall judgment.

3. **Expected Evidence Distribution:** Typically, a claim should have about 3 supporting evidence pieces, with the remainder being a mix of neutral and refuting evidence if available.

4. **Reason Field:** Fill 'Reason' field with text that explains the assigned label. This can be a novel explanation or drawn directly from the evidence.

A.1.3 Data Consolidation and Preprocessing Steps

- *Text Cleaning & Preprocessing*: Extraneous characters and irrelevant columns were removed to ensure consistency. Temporary columns, such as the 'Old Reason' column left by different annotation teams, were dynamically updated or removed as needed.
- *Claim & Evidence ID Generation*: Unique Claim_Topic_ID and Evidence_Topic_IDs were generated to maintain traceability between claims and their corresponding evidence, as detailed in Table 7.1.
- *Annotation Tracking & Traceability*: After cleaning the individual files from each annotation team, the Annotation_ID column was added to the dataset, with the same ID assigned to all instances curated by the respective team (B_2.0, C_2.1, C_2.2), enabling traceability back to the raw files.
- *Dataset Consolidation*: The cleaned datasets were concatenated into a single DataFrame, and the index was reset to ensure consistency and avoid any indexing issues across the combined dataset.
- *Consistency in Labeling & Reason Propagation*: Claims with the same Claim_Topic_ID were assigned the same propagated human-generated reasons to ensure consistency within each claim group, maintaining coherence across the related evidence pieces.
- *Validation & Reason Type Assignment*: The dataset was validated by grouping entries by Claim_Topic_ID to verify that each claim was correctly linked with its evidence and explanation. The reasons were further categorized into Abstractive and Extractive, as described in Table 7.1.

A.2 SOI Generation Algorithm

Algorithm 2 provides the complete process for generating the SOI (discussed in Chapter 5), covering data filtering, embedding generation, clustering, and the similarity-based selection of relevant evidence and claims.

Algorithm 2 SOI Generation with Thematic Clustering and Filtering

Require: Dataset D, Selected Claim ID c_{selected}, Similarity Threshold δ, Number of Clusters k

Ensure: SOI (Subset of Interest) for the selected claim

Data Preparation
Load dataset D
Filter dataset to obtain thematic subset D_T based on the selected theme T
for each claim c_i in D_T **do**
    Extract associated evidences $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,m}\}$
end for
Embedding Generation
for each claim c_i in D_T **do**
    Generate embedding $\text{emb}_{c_i} = \text{Transformer}(c_i)$
end for
for each evidence $e_{i,j} \in E_i$ for each claim c_i in D_T **do**
    Generate embedding $\text{emb}_{e_{i,j}} = \text{Transformer}(e_{i,j})$
end for
Thematic Clustering
Aggregate all embeddings $\text{emb}(D_T) = \{\text{emb}_{c_i}, \text{emb}_{e_{i,j}} \mid c_i \in D_T, e_{i,j} \in E_i\}$
Apply Gaussian Mixture Model (GMM) with Expectation-Maximization (EM) to cluster the aggregated embeddings
$L = \text{EM}(\text{emb}(D_T)) = \arg\max_\theta \sum_{i=1}^{n} \log p(\text{emb}_t \mid \theta)$
Obtain cluster labels L for each claim and evidence based on their embeddings
Cluster Filtering Based on Selected Claim
Determine the cluster C_{selected} that the selected claim c_{selected} belongs to
Extract all claims and evidences belonging to the selected cluster C_{selected}
SOI Generation
Initialize empty SOI dictionary: $\text{SOI} = \{\}$
Add claim c_{selected} to SOI
Add its annotated evidences to SOI
for each related claim c_k in C_{selected} **do**
    if $\text{sim}(c_k, c_{\text{selected}}) > \delta$ **then**
        Add c_k to SOI
        for each evidence e_k associated with c_k **do**
            if $\text{sim}(e_k, c_{\text{selected}}) > \delta$ **then**
                Add e_k to SOI
            end if
        end for
    end if
end for
Output: Return the comprehensive SOI for the selected claim, including all relevant evidences and related claims exceeding the similarity threshold.

Appendix B

B.1 Retrieved Evidence of the Case Study

The retrieved evidence by CARAG-u for the claim "Climate change is caused by atmospheric concentration of greenhouse gases,", is presented, categorized under different α configurations. The results demonstrate that thematic alignment improves as α decreases. The post-hoc explanations featured in Table 9.2 in the case study (Section 9.3.2) are generated based on these evidence sets.

- $\alpha = 1, \delta = 0.75$
  They could drive back to Israel and get a ferry around the Middle East to Qatar - but that is a two-week journey which would have meant missing Wales' first match. Farley explained that this will no longer be the case in the future. In January 2022, Brookings Metro published a report that assessed the impact of long COVID on the labor market. Data on the condition's prevalence was limited, so the report used various studies to make a conservative estimate: 1.6 million full-time equivalent workers could be out of work due to long COVID. Scientists warn that it is likely better at dodging immunity than others based on lab tests of its grandfather BA.2.75.

- $\alpha = 0.75, \delta = 0.75$
  They could drive back to Israel and get a ferry around the Middle East to Qatar - but that is a two-week journey which would have meant missing Wales' first match. Farley explained that this will no longer be the case in the future. In January 2022, Brookings Metro published a report that assessed the impact of long COVID on the labor market. Data on the condition's prevalence was limited, so the report used various studies to make a conservative estimate: 1.6 million full-time equivalent workers could be out of work due to long COVID. Of those, 2 to 4 million are out of work due to long COVID.

- $\alpha = 0.5, \delta = 0.75$
  They could drive back to Israel and get a ferry around the Middle East to Qatar - but that is a two-week journey which would have meant missing Wales' first match. In January 2022, Brookings Metro published a report that assessed the impact of long COVID on the labor market. Data on the condition's prevalence was limited, so the report used various studies to make a conservative estimate: 1.6 million full-time equivalent workers could be out of work due to long COVID. Farley explained that this will no longer be the case in the future.

- $\alpha = 0.25, \delta = 0.75$
  There are a number of industrial decarbonization schemes, including energy efficiency and emissions trading schemes, that we administer for the UK and regulate in England. At the heart of all these schemes is the goal of reducing greenhouse gas emissions with a particular focus on carbon emissions. This greenhouse gas trading scheme forms part of the UK government's ambition to achieve net zero emissions by 2050. Heavy-duty, diesel-powered transportation is considered one of the largest contributors to greenhouse gas emissions (GHG), but by accelerating the path toward a ZE-MHDV market, communities around the world will benefit from cleaner air, better health, and technology innovation to help bolster their economies.

Appendix C

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
AFV	Automated Fact Verification
ELF	Explanation Learn-Friendly
DNN	Deep Neural Network
RTE	Recognizing Textual Entailment
GMM	Gaussian Mixture Models
EM	Expectation-Maximization
SOI	Subset of Interest
RAG	Retrieval-Augmented Generation