applied sciences

MDPI

Article

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

Manju Vallayil [1,*,†], Parma Nand [1,†], Wei Qi Yan [1,†], Héctor Allende-Cid [2,3,4] and Thamilini Vamathevan [5]

[1] School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand; parma.nand@aut.ac.nz (P.N.); weiqi.yan@aut.ac.nz (W.Q.Y.)
[2] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340025, Chile; hector.allende@pucv.cl
[3] Knowledge Discovery Department, Fraunhofer-Institute of Intelligent Analysis and Information Systems (IAIS), 53757 Sankt Augustin, Germany
[4] Lamarr Institute for Machine Learning and Artificial Intelligence, 53115 Dortmund, Germany
[5] Tureya Limited, Auckland 1024, New Zealand; thamil@tureya.co.nz
* Correspondence: manju.vallayil.vijayalekshmi@aut.ac.nz
† Current address: Institute of Robotics and Vision (IoRV), WZ801A WZ Building, Level 8, 6-24 St Paul St, Auckland 1010, New Zealand.

Abstract: This study introduces an explainable framework for Automated Fact Verification (AFV) systems, integrating a novel Context-Aware Retrieval and Explanation Generation (CARAG) methodology. CARAG enhances evidence retrieval by leveraging thematic embeddings derived from a Subset of Interest (SOI, a focused subset of the fact-verification dataset) to integrate local and global perspectives. The retrieval process combines these thematic embeddings with claim-specific vectors to refine evidence selection. Retrieved evidence is integrated into an explanation-generation pipeline employing a Large Language Model (LLM) in a zero-shot paradigm, ensuring alignment with topic-based thematic contexts. The SOI and its derived thematic embeddings, supported by a visualized SOI graph, provide transparency into the retrieval process and promote explainability in AI by outlining evidence-selection rationale. CARAG is evaluated using FactVer, a novel explanation-focused dataset curated to enhance AFV transparency. Comparative analysis with standard Retrieval-Augmented Generation (RAG) demonstrates CARAG's effectiveness in generating contextually aligned explanations, underscoring its potential to advance explainable AFV frameworks.

Keywords: explainable AI; XAI; automated fact verification; AFV; global explainability; RAG

Academic Editors: Luca Longo, Mario Brcic and Sebastian Lapuschkin

Received: 31 December 2024
Revised: 12 February 2025
Accepted: 12 February 2025
Published: 13 February 2025

Citation: Vallayil, M.; Nand, P.; Yan, W.Q.; Allende-Cid, H.; Vamathevan, T. CARAG: A Context-Aware Retrieval Framework for Fact Verification, Integrating Local and Global Perspectives of Explainable AI. *Appl. Sci.* **2025**, *15*, 1970. https://doi.org/10.3390/app15041970

1. Introduction

Explainability, defined as the ability to interpret model behavior in a human-understandable way [1], is increasingly essential in AI applications such as Automated Fact Verification (AFV) systems. While recent advancements in AI architectures, such as transformer models [2] and Retrieval-Augmented Generation (RAG) [3], 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 [4]. This architectural

approach was later adopted by several subsequent researchers such as [4–10]. This architecture involves the composite tasks of collecting or retrieving relevant evidence to support or refute a claim, ranking these pieces of evidence by importance, 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 [11]. 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) [12–15]. 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 highlights the need for greater transparency in AI systems [12,13,16,17].

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 [18,19] risk producing incomplete or misleading explanations. Logic-based models like LOREN [20] and ProoFVer [10] create transparent explanations through logic rules but are difficult to scale and apply in real-world contexts. Attention mechanisms [21–23] highlight important features, yet their reliability is questionable as attention scores do not consistently align with key decision-making features [11]. Counterfactual explanations [24,25] 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 [10], 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 [15] and the extensive survey on explainable automated fact-checking by [26], 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 [27] and MultiFC [28] 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 [29]. 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 [26]. 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 [30]. Different

researchers interpret and apply these concepts inconsistently [1,31–33], 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 [22], while others conflate interpretability with explainability, restricting it to individual explanations [18]. 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 2, we describe the dataset in Section 3 and the methodology in Section 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 5, while challenges, future research directions, and conclusions are discussed in Sections 6–8, 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.

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 [34]. 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 [35,36]. 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 [4]. While Wikipedia offered broad accessibility, it also introduced challenges with comprehensiveness and potential biases, impacting the fidelity of resulting AFV models [37]. 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 [38]. Nonetheless, interpretability, accuracy, and fidelity remain essential paradigms in explainable AI, as emphasized by recent work on XAI [1,13,31].

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 [38] 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 [39] and SHAP values [40], as well as surrogate models like LIME [41], 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 [42] and RoBERTa [43], 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 [44] and Llama 3 [45], which respond based on pre-trained knowledge; or as with assistant models like GPT-4 by OpenAI [46], Claude by Anthropic [47], which undergo additional alignment through methods like instruction tuning and Reinforcement Learning from Human Feedback (RLHF) to perform user-specific tasks [30]. The prompting paradigm is further reinforced by the impressive zero-shot performance of LLMs in various language tasks [48], 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 [30].

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 4 details our framework, addressing the roles, paradigms, and scope of XAI in AFV comprehensively for a balanced explainability.

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 [15], 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.

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 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, describes the preparation process, and presents example data entries to illustrate its structure and composition.

Table 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.

3.2. Dataset Description

The dataset includes 589 unique claims. As shown in Figure 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.

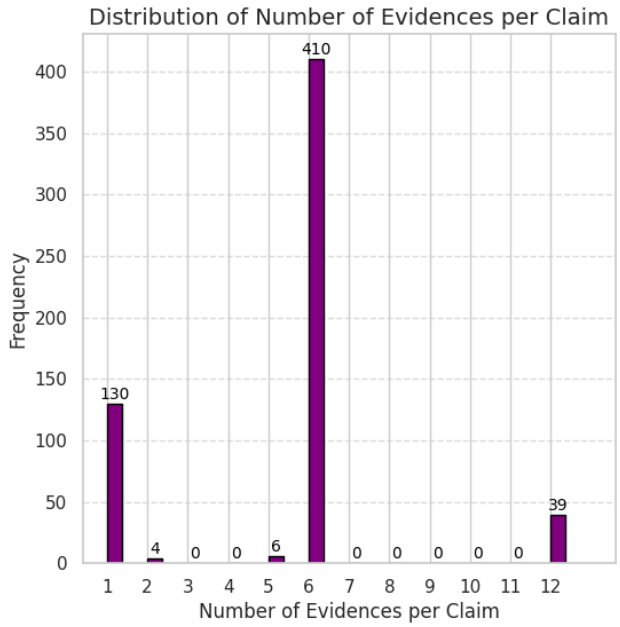


Figure 1. Distribution of the amount of evidence associated with each claim across all themes. The figure illustrates that most claims are supported by six pieces of evidence (represented by the tallest bar), with smaller subsets having either one or twelve pieces of evidence, highlighting the variability in evidence distribution across the dataset.

Furthermore, Figure 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 to 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.

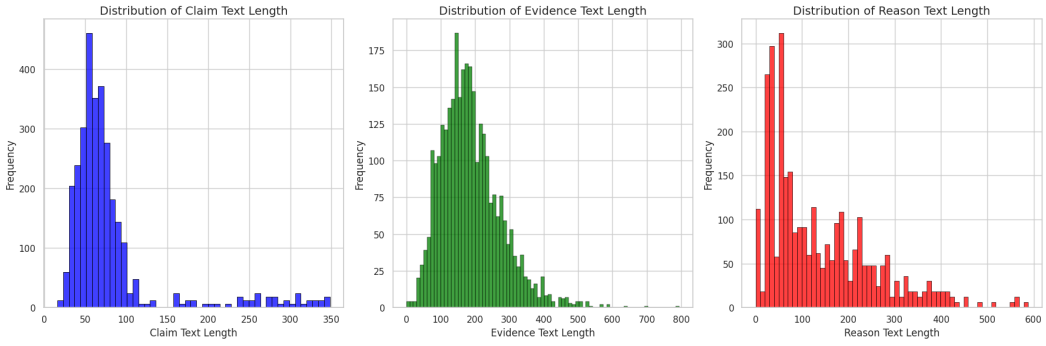


Figure 2. Distribution of claim, evidence, and reason text length.

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.

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 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. 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 3.2 and represented in Figure 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). This consolidated dataset, named FactVer_1.3, is designed to facilitate AFV and support XAI research in this domain.

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 Item 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 Item 2: As of July, 46 countries are pricing emissions through carbon taxes or emissions trading schemes (ETSs), according to the International Monetary Fund.
  - Evidence Item 3: NZ's agricultural emissions are not currently captured under the ETS (unlike other sources like industrial processes).
  - Evidence Item 4: The number of emission units released for auction is designed to meet New Zealand's international obligations.

  - Evidence Item 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 Item 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

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 (dataset available at: https://huggingface.co/datasets/manjuvallayil/factver_master, accessed on 31 December 2024), and the associated code is available on GitHub (code repository available at: https://github.com/manjuvallayil/factver_dev, accessed on 31 December 2024).

4. Methodology

This section presents our Context-Aware 'Retrieval Augmented Generation' Framework (CARAG), an approach to enhancing 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 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 on in the following sections.

Figure 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.

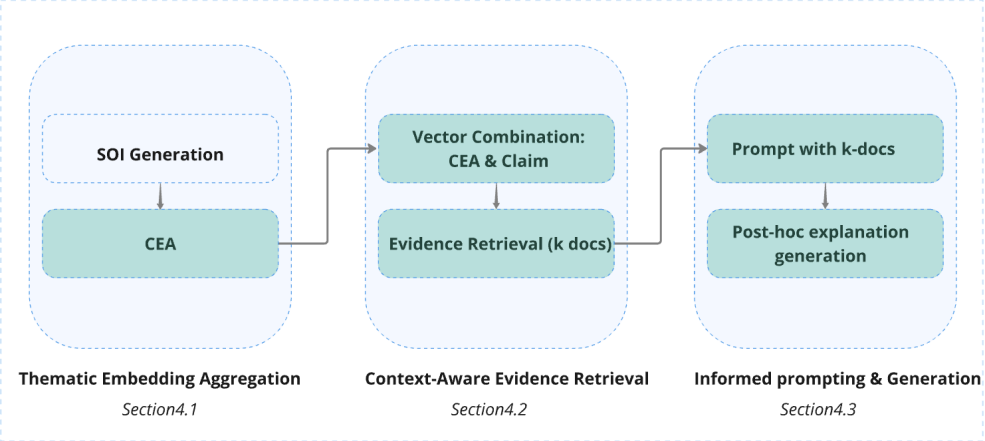


Figure 3. Overview of the methodology components.

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 [49]. 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 on in Section 5.1.1.

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 the following: (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 hereafter referred to as thematic cluster evidence). Importantly, not all annotated evidence of the selected claim or related claims will necessarily be present in the identified cluster, as it is the result of an unsupervised clustering process.

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 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 item or claim to the selected claim. (The complete algorithm for SOI generation is provided in Appendix B for those interested in the technical details).

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

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 [49], 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.

Table 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.

To implement this SOI integration into CARAG, selected elements from the SOI dictionary are extracted, specifically, 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. Next, as represented in Equation (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 [50–52].

In our pipeline, we define this phase as Contextual Embedding Aggregation (CEA), and Figure 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.

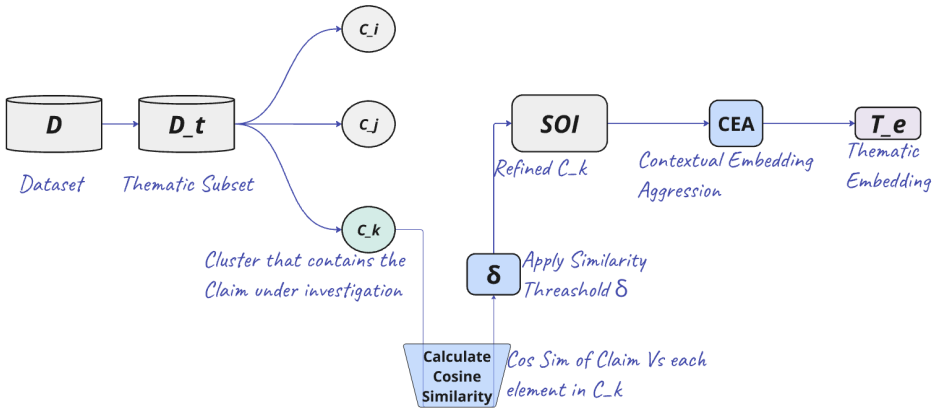


Figure 4. The figure illustrates the curation of the unified thematic embedding (*T_e*) process from the fact-verification dataset (*D*). The claim's cluster (*C_k*), identified from the thematic subset (*D_t*), is refined using cosine similarity and a similarity threshold (δ) to form the Subset of Interest (*SOI*) for the claim. This SOI is then processed through Contextual Embedding Aggregation (CEA) using Equation (1) to generate the final thematic embedding.

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 (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 4.2).

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 retrieving evidence from the (vectorized) fact-verification database. These embeddings are merged using a weighted mechanism (Equation (2)), where the parameter α controls the balance between claim-specific details and thematic context.

$$\text{Combined\_Embedding} = \alpha \cdot \text{Claim\_Embedding} + (1 - \alpha) \cdot \text{Thematic\_Embedding} \quad (2)$$

As established in Section 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.

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.

Figure 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 (1), and combines it with the claim vector using a weighted averaging process (Equation (2)). This weighted approach produces a more refined final combined embedding used for querying (represented by the purple dots in Figure 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 items, and specific instructions. The LLM (e.g., LLaMA) subsequently generates a concise explanation, assessing the claim's veracity and offering a justification.

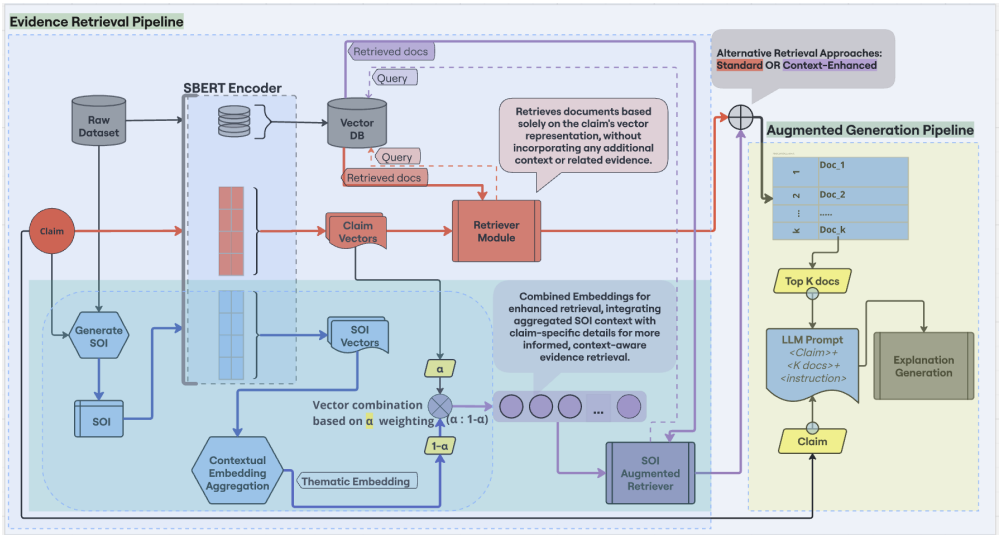


Figure 5. Overview of the CARAG Framework vs. Standard RAG. The Standard Retrieval path (red arrows) retrieves evidence solely based on the claim vector, without incorporating contextual information. In contrast, the CARAG pipeline (light blue-shaded area) introduces context-aware evidence retrieval, 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.

5. Experimental Framework and Results

This section describes the key elements employed in our experimental framework and outlines a case study comparison of explanation approaches (Section 5.1) alongside a comparative analysis of RAG and CARAG methods (Section 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.

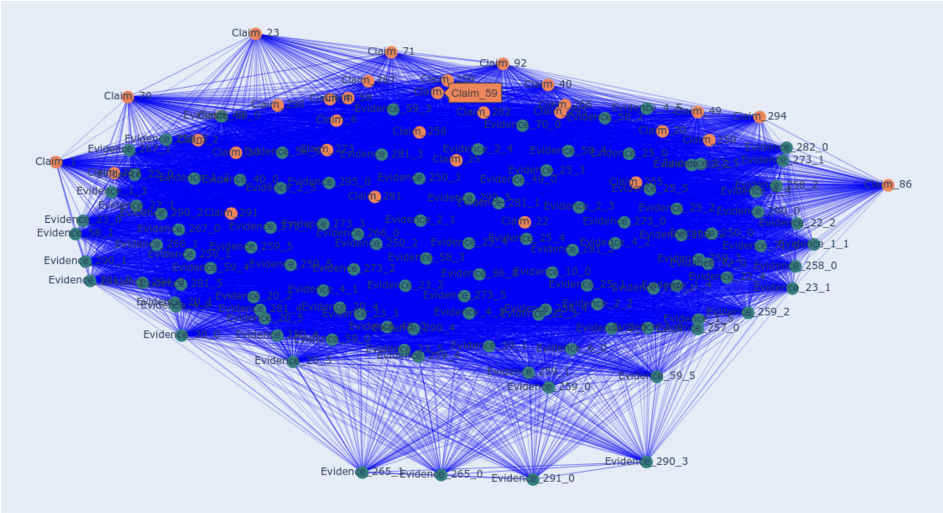
For embedding generation, we selected the Sentence-BERT (SBERT) model [53], which enhances BERT by incorporating siamese and triplet network structures to produce semantically meaningful sentence vectors. Specifically, we employed the open-source *all-mpnet-base-v2* variant of SBERT, fine-tuned on over 1 billion textual pairs. This model is relevant in our methodology, where cosine similarity supports context filtering and nuanced textual similarity [54]. Moreover, we chose to use the same SBERT encoder as employed in the SOI methodology [49] 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 [55] 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.

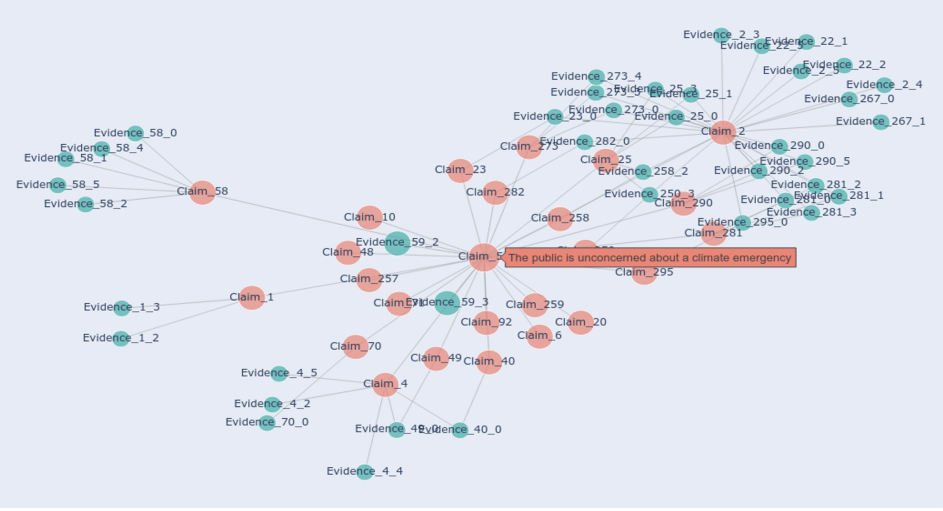
For fact verification and explanation generation, we employed the *Llama-2-7b-chat-hf* variant of LLaMA from Meta [56], chosen for its balance of efficiency and performance and its compatibility with our computational resources. With 7 billion parameters, Llama-2 Chat is suited to our explainability tasks, offering competitive performance comparable to models like ChatGPT and PaLM [56]. Optimized for dialogue and trained with RLHF, the model supports our informed prompting methodology (Section 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. An example of this workflow is detailed in Section 5.1. To ensure unbiased results, GPU memory is cleared before each generation run. Future work could explore newer versions, such as Llama 3 [45], on advanced hardware to assess potential improvements.

To visualize the thematic clusters and SOI of a claim, as shown in Figure 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 5.2.

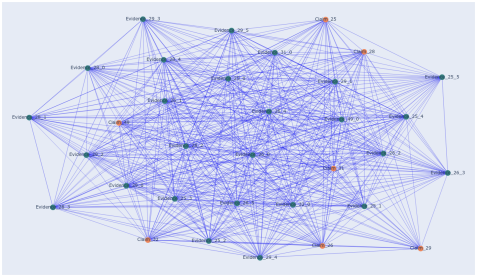


(**a**) Cluster 1: Full thematic cluster containing Claim 59

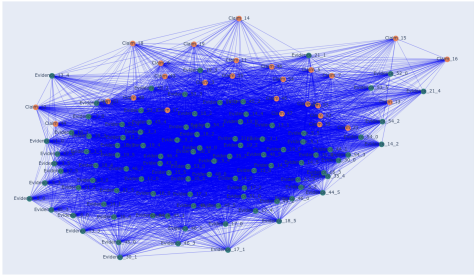


(**b**) SOI for Claim 59: Subset derived from Cluster 1

Figure 6. *Cont.*



(**c**) Cluster 0: Secondary thematic cluster for comparison



(**d**) Cluster 2: Tertiary thematic cluster for comparison

Figure 6. Visualization of SOI and thematic clusters within the Climate theme. (**a**,**c**,**d**) The identified clusters (Clusters 1, 0, and 2, respectively). In (**a**), Cluster 1 highlights Claim 59 for clarity, while (**b**) shows the refined SOI for Claim 59 derived from Cluster 1.

This architecture enables streamlined integration of the AFV pipeline, supporting both RAG and CARAG methods for a comprehensive analysis, as shown in Figure 5 and empirically discussed in subsequent sections as we evaluate the framework using instances from our fact-verification dataset FactVer, introduced in Section 3.

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 4.1, 4.2 and 4.3, respectively, illustrating the practical application of our structured methodology for Claim 59.

5.1.1. Thematic Embedding Generation for Claim 59

To generate a thematic embedding for Claim 59 from *FactVer* using CEA (Section 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 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 [57,58] 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 [57,58]. 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 item to a single cluster to ensure clear relationships and facilitate precise analysis in AFV [49]. Claim 59 is identified within Cluster 1, a dense network containing 85 nodes and 3103 edges, indicating a rich interconnection of semantically related claims and evidence. Following the methodology outlined in Section 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 2, providing a structured foundation for embedding generation.

Figure 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 items (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 (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.

5.1.2. 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 (2), as detailed in Section 4.2); and (3) using CARAG with a balanced combination of the claim vector and thematic embedding by setting $\alpha = 0.5$ in Equation (2).

For each approach, we selected the top $k = 6$ evidence items, in alignment with our dataset distribution statistics (Section 3.2), which indicate that the majority of claims are supported by six pieces of evidence. Table 3 presents a side-by-side comparison of evidence retrieved by these three approaches for Claim 59.

Table 3. Comparison of annotated evidence and 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.	1. Failure will result in the country's once-successful car-making industry being scrapped.	1. Failure will result in the country's once-successful car-making industry being scrapped.
2. U.N. notes 1.5 degrees Celsius as a crucial limit.	2. Two-week journey which would have meant missing Wales' first match.	2. Greenhouse gas trading scheme forms part of the UK government's ambition to achieve net-zero emissions by 2050.
3. Todd's opposition to increasing traffic during a climate emergency.	3. Shares fell as low as $6.50-apiece on Monday, down 97 percent from August 2021.	3. Short-notice public investigatory attention affecting businesses.
4. Declaration of a climate emergency in 2020.	4. BMW-branded cars, motorcycles, and Mini models sold since October 1 get the new warranty.	4. Korean officials discussing options to correct unfair EV tax policies.
5. Groundswell NZ's protest against He Waka Eke Noa.	5. Korean officials discussing several possible options to correct what they believe to be unfair policies that eliminated up to $7500 of tax credits for EVs produced outside North America.	5. Two-week journey potentially missing Wales' first match.
6. Chinese protests demanding action on climate change.	6. White House statement about Biden's health condition.	6. Reliance on imports from China, US, and Europe for the car industry.

A key observation from the evidence comparison in Table 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.

5.1.3. 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 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>`
> (An example for specific instruction is, 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))

`<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 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 3, respectively.

Figure 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.

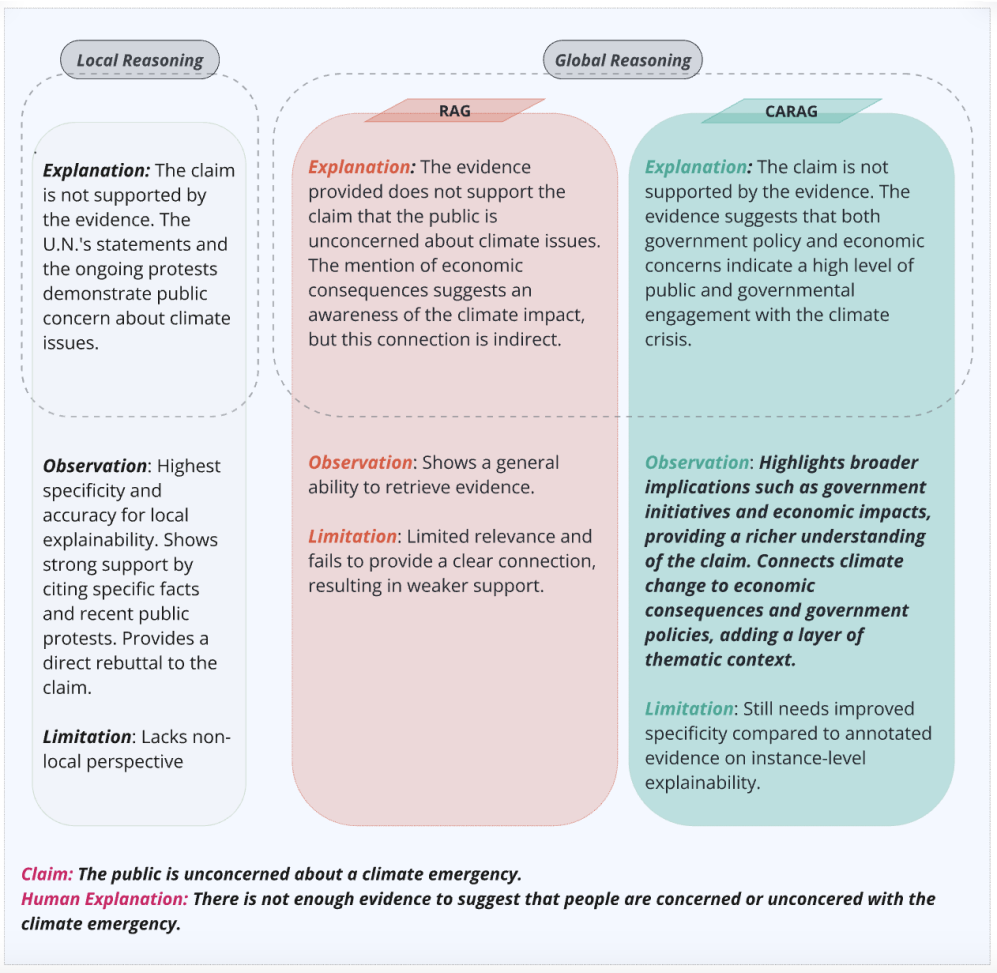


Figure 7. Comparison of generated explanations across retrieval approaches for Claim 59, showcasing the methodology's application.

The qualitative comparison in Figure 7 further classifies the explanations into local and global reasoning. Explanations based on annotated evidence (left) provide 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.

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 [59]. As illustrated in Figure 6, Panel (b), visual explanations use graphical elements to clarify decision-making processes, while Figure 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.

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 (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.

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 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.

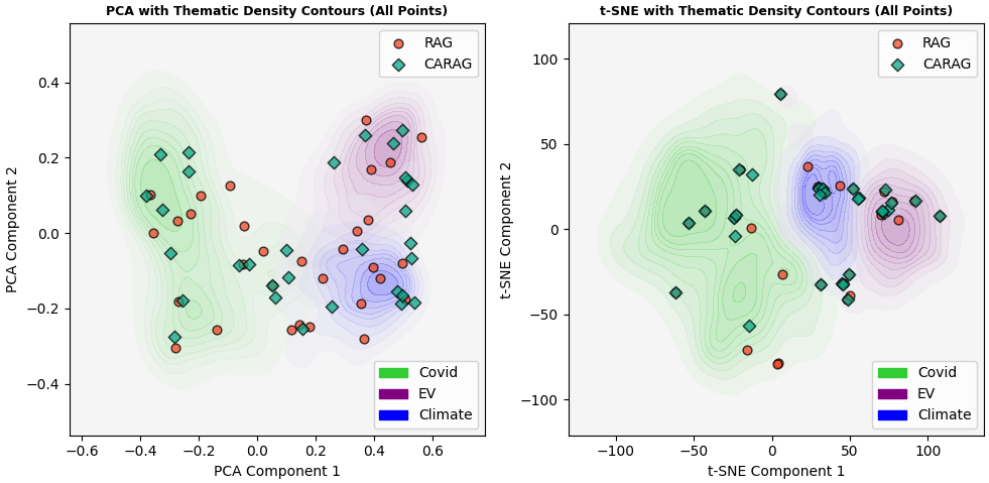
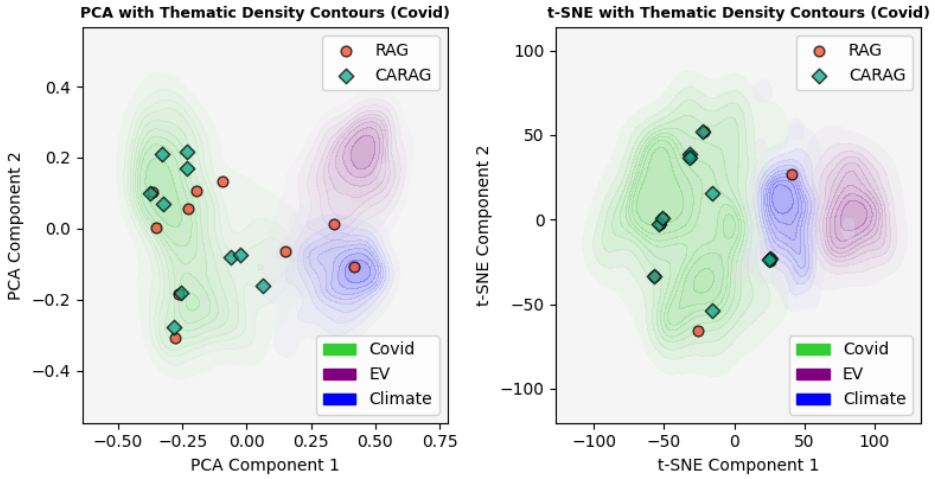


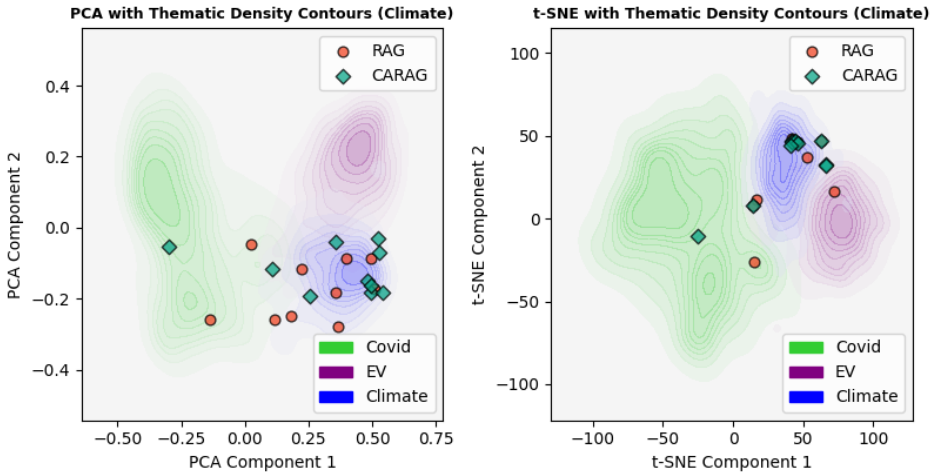
Figure 8. PCA (**left**) and t-SNE (**right**) visualizations of embedding distributions for RAG-generated explanations (red circles) and CARAG-generated explanations (green diamonds), shown with KDE-based thematic density contours in the background (green for COVID, blue for climate, and purple for electric vehicles). These contours illustrate thematic boundaries, enabling a comparative evaluation.

To provide more granular insights into each theme, we present separate plots for each theme in Figure 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 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.

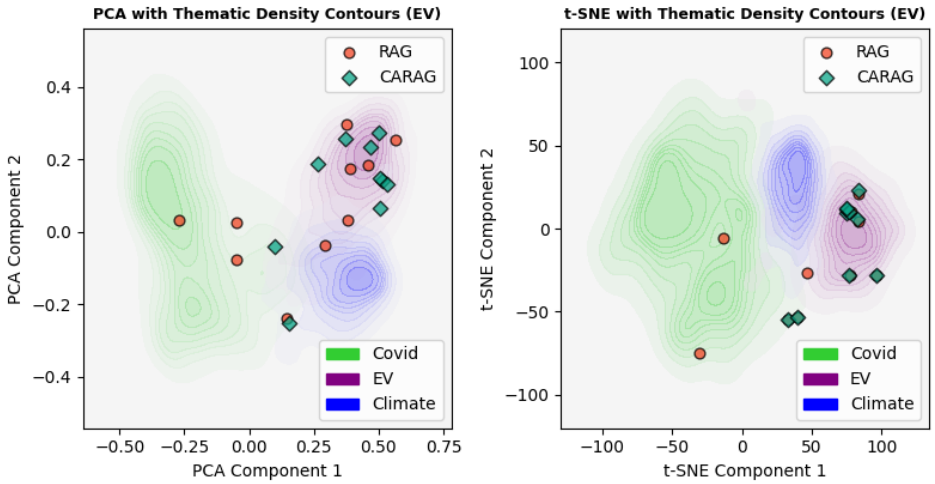
RAG's distribution reveals a tendency to capture generalized information across themes, which aligns with its retrieval-augmented nature but may dilute thematic specificity. Conversely, CARAG's thematic retrieval is more focused, producing explanations that closely align with each theme's contours. By 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.



(**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

Figure 9. PCA (**left**) and t-SNE (**right**) visualizations of embedding distributions for RAG-generated explanations (red circles) and CARAG-generated explanations (green diamonds) for claims related to COVID, climate, and electric vehicles, shown in (**a**–**c**), respectively. Each panel includes KDE-based thematic density contours in the background (green for COVID, blue for climate, and purple for electric vehicles), highlighting each approach's alignment with underlying thematic regions.

The quantitative results in Table 4 corroborate the visual patterns observed in Figure 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 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 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.

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.

Table 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**

Table 4. *Cont.*

(**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**

Table 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.

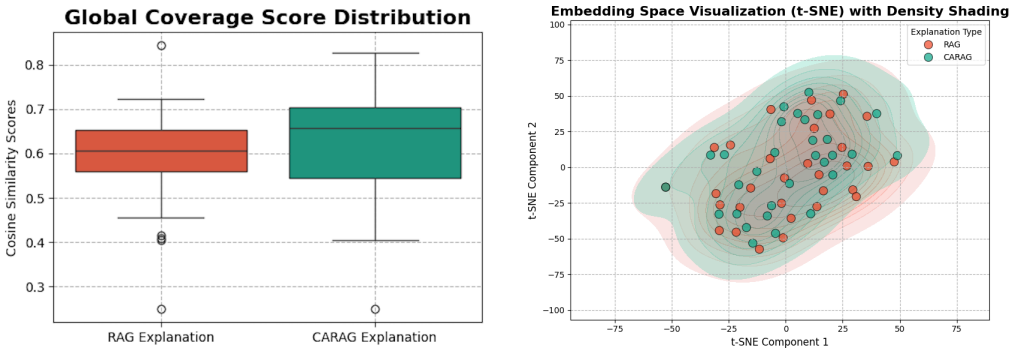
Theme	RAG (PCA) Avg	CARAG (PCA) Avg	Diff (PCA) Avg	RAG (t-SNE) Avg	CARAG (t-SNE) Avg	Diff (t-SNE) Avg
Covid	0.2732	0.2120	**−0.0612**	31.6718	31.5578	**−0.1140**
Climate	0.2056	0.2094	**0.0037**	37.4171	37.3071	**−0.1100**
Electric Vehicles	0.2881	0.1858	**−0.1023**	76.4661	76.2973	**−0.1687**
Overall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

5.3. Limitations of Standard Analysis and 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, 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, 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 10a 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 10b, highlights clustering within the embedding space without indicating clear thematic boundaries. Unlike our PCA and t-SNE approach in Section 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

Figure 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 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.

6. Challenges and Limitations

In AFV, explainability can be approached from three primary perspectives, architectural, methodological, and data [15]. 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 [30], 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 predefined themes. These updates will enable CARAG to generalize across diverse datasets, with preliminary results to be presented in an upcoming publication.

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 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.

8. Conclusions

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 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.

Author Contributions: Conceptualization, M.V. and P.N.; methodology, M.V.; software, M.V.; validation, M.V.; formal analysis, M.V.; investigation, M.V.; resources, M.V. and P.N.; data collection, M.V.; writing—original draft preparation, M.V.; writing—review and editing, M.V., P.N., W.Q.Y., H.A.-C. and T.V.; visualization, M.V.; supervision, P.N., W.Q.Y. and H.A.-C.; project administration P.N. and W.Q.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated and used in this study [https://huggingface.co/datasets/manjuvallayil/factver_master] (accessed on 31 December 2024).

Acknowledgments: The authors greatly appreciate the contributions of the three annotation teams, who worked diligently. Their efforts were pivotal in creating the datasets that form the foundation of this research. In addition, the authors wish to thank the School of Engineering, Computer and Mathematical Sciences (ECMS) at Auckland University of Technology for providing access to GPU resources, which were instrumental in enabling the computational work for this research.

Conflicts of Interest: Author Thamilini Vamathevan was employed by the company Tureya Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
AFV	Automated Fact Verification
RAG	Retrieval-Augmented Generation
GMM	Gaussian Mixture Models
EM	Expectation-Maximization
SOI	Subset of Interest

Appendix A. Template and Instructions for Annotation

Appendix A.1. Steps for Annotation

1. Obtain the template: Obtain the Excel template file containing fields as specified in Table A1.
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 six rows for each claim, each containing a different evidence span. If no additional evidence is found, fill up with neutral text to reach six 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.

Table A1. 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 item, ranging from E1 to E6, to ensure traceability between claims and their corresponding evidence item.
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.

Appendix A.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 six 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 three supporting evidence pieces, with the remainder being a mix of neutral and refuting evidence if available.
4. Reason field: Fill the 'Reason' field with text that explains the assigned label. This can be a novel explanation or drawn directly from the evidence.

Appendix A.3. Data Consolidation and Preprocessing Steps

- Text cleaning and 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 and 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 1.
- Annotation tracking and 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 and 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 and 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 1.

Appendix B. SOI Generation

Algorithm A1 provides the complete process for generating the SOI, covering data filtering, embedding generation, clustering, and the similarity-based selection of relevant evidence and claims.

Algorithm A1 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 evidence items $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,m}\}$
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 item $e_{i,j} \in E_i$ for each claim c_i in D_T **do**
    Generate embedding $\text{emb}_{e_{i,j}} = \text{Transformer}(e_{i,j})$
end for

Algorithm A1 *Cont.*

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 evidence items belonging to the selected cluster C_{selected}
SOI Generation
Initialize empty SOI dictionary: $\text{SOI} = \{\}$
Add claim c_{selected} to SOI
Add its annotated evidence items 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 item 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 evidence items and related claims exceeding the similarity threshold.

References

1. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Vollume 2017, pp. 5999–6009.
3. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kwiatkowski, T.; Parikh, A.; Boyd-Graber, J.; Riedel, S. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20), Vancouver, BC, Canada, 6–12 December 2020.
4. Thorne, J.; Vlachos, A.; Cocarascu, O.; Christodoulopoulos, C.; Mittal, A. The Fact Extraction and VERification (FEVER) Shared Task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium, 1 November 2018; pp. 1–9. [CrossRef]
5. Soleimani, A.; Monz, C.; Worring, M. BERT for evidence retrieval and claim verification. *arXiv* **2019**, arXiv:1910.02655.
6. Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; Yin, J. Reasoning over semantic-level graph for fact checking. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6170–6180. [CrossRef]
7. Jiang, K.; Pradeep, R.; Lin, J. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In Proceedings of the ACL-IJCNLP 2021—59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Vollume 2, pp. 402–410. [CrossRef]
8. Chen, J.; Zhang, R.; Guo, J.; Fan, Y.; Cheng, X. GERE: Generative Evidence Retrieval for Fact Verification. In Proceedings of the SIGIR 2022—Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 2184–2189. [CrossRef]
9. DeHaven, M.; Scott, S. BEVERS: A General, Simple, and Performant Framework for Automatic Fact Verification. In Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), Dubrovnik, Croatia, 5 May 2023; pp. 58–65.
10. Krishna, A.; Riedel, S.; Vlachos, A. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1013–1030.
11. Guo, Z.; Schlichtkrull, M.; Vlachos, A. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 178–206. [CrossRef]

12. Gunning, D.; Vorm, E.; Wang, J.Y.; Turek, M. DARPA's explainable AI (XAI) program: A retrospective. *Appl. AI Lett.* **2021**, *2*, e61. [CrossRef]
13. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805.
14. Kim, T.W. Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test. *arXiv* **2018**, arXiv:1810.09598.
15. Vallayil, M.; Nand, P.; Yan, W.Q.; Allende-Cid, H. Explainability of automated fact verification systems: A comprehensive review. *Appl. Sci.* **2023**, *13*, 12608. [CrossRef]
16. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
17. Gunning, D. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*; Technical Report; Defense Advanced Research Projects Agency: Arlington, VA, USA, 2016.
18. Atanasova, P.; Simonsen, J.G.; Lioma, C.; Augenstein, I. Generating Fact Checking Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7352–7364. [CrossRef]
19. Kotonya, N.; Toni, F. Explainable automated fact-checking for public health claims. *arXiv* **2020**, arXiv:2010.09926. [CrossRef]
20. Chen, J.; Bao, Q.; Sun, C.; Zhang, X.; Chen, J.; Zhou, H.; Xiao, Y.; Li, L. Loren: Logic-regularized reasoning for interpretable fact verification. In Proceedings of the AAAI Conference on Artificial Intelligence, Pomona, CA, USA, 24–28 October 2022; Volume 36, pp. 10482–10491.
21. Popat, K.; Mukherjee, S.; Yates, A.; Weikum, G. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv* **2018**, arXiv:1809.06416.
22. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.
23. Amjad, H.; Ashraf, M.S.; Sherazi, S.Z.A.; Khan, S.; Fraz, M.M.; Hameed, T.; Bukhari, S.A.C. Attention-Based Explainability Approaches in Healthcare Natural Language Processing. In Proceedings of the International Conference on Health Informatics (HEALTHINF), Kyoto, Japan, 12–14 May 2023; pp. 689–696.
24. Dai, S.C.; Hsu, Y.L.; Xiong, A.; Ku, L.W. Ask to Know More: Generating Counterfactual Explanations for Fake Claims. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2022; pp. 2800–2810.
25. Xu, W.; Liu, Q.; Wu, S.; Wang, L. Counterfactual Debiasing for Fact Verification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 6777–6789.
26. Kotonya, N.; Toni, F. Explainable Automated Fact-Checking: A Survey. *arXiv* **2020**, arXiv:2011.03870.
27. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. 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, New Orleans, LA, USA, 1–6 June 2018; pp. 809–819. [CrossRef]
28. Augenstein, I.; Lioma, C.; Wang, D.; Chaves Lima, L.; Hansen, C.; Hansen, C.; Simonsen, J.G. 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), Hong Kong, China, 3–7 November 2019; pp. 4685–4697. [CrossRef]
29. Stammbach, D.; Ash, E. e-FEVER: Explanations and Summaries for Automated Fact Checking. In Proceedings of the Conference for Truth and Trust Online (TTO 2020) (Virtual), Online, 16–17 October 2020. [CrossRef]
30. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–38. [CrossRef]
31. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
32. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef] [PubMed]
33. Moradi, M.; Samwald, M. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Syst. Appl.* **2021**, *165*, 113941.
34. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. 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, Copenhagen, Denmark, 9–11 September 2017; pp. 2931–2937. [CrossRef]
35. Shi, B.; Weninger, T. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowl.-Based Syst.* **2016**, *104*, 123–133. [CrossRef]

36. Gardner, M.; Mitchell, T. Efficient and expressive knowledge base completion using subgraph feature extraction. In Proceedings of the Conference Proceedings—EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1488–1498. [CrossRef]
37. Hanselowski, A.; Stab, C.; Schulz, C.; Li, Z.; Gurevych, I. A richly annotated corpus for different tasks in automated fact-checking. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 493–503. [CrossRef]
38. Singhal, R.; Patwa, P.; Patwa, P.; Chadha, A.; Das, A. 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), Miami, FL, USA, 15 November 2024.
39. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
40. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume. 30.
41. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
42. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2–7 June 2019.
43. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
44. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
45. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
46. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712.
47. AnthropicAI. Introducing Claude. 2023. Available online: https://www.anthropic.com/index/introducing-claude (accessed on 31 December 2024).
48. Liusie, A.; Manakul, P.; Gales, M.J.F. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. *arXiv* **2024**, arXiv:2307.07889.
49. Vallayil, M.; Nand, P.; Yan, W.Q. Explainable AI through Thematic Clustering and Contextual Visualization: Advancing Macro-Level Explainability in AFV Systems. In Proceedings of the ACIS 2024 Proceedings, Canberra, Australia, 4–6 December 2024; Number 101.
50. Zhao, W.; Zheng, W.; Wang, Y.; Wang, T. Fake News Detection Based on Knowledge-Guided Semantic Analysis. *Electronics* **2024**, *13*, 259. [CrossRef]
51. Tang, J.; Yin, D.; Zhang, J.; Yin, J. Secure Embedding Aggregation for Federated Representation Learning. In Proceedings of the 2023 IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan , 25–30 June 2023; pp. 2392–2397.
52. Iliadis, D.; Peikou, M.; Adamidou, C.; Kyriakopoulou, A. A comparison of embedding aggregation strategies in drug–target interaction prediction. *BMC Bioinform.* **2024**, *25*, 59. [CrossRef] [PubMed]
53. Reimers, N.; Gurevych, I. 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), Hong Kong, China, 3–7 November 2019.
54. Jayanthi, S.M.; Embar, V.; Raghunathan, K. Evaluating Pretrained Transformer Models for Entity Linking in Task-Oriented Dialog. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), Silchar, India, 16–19 December 2021; Bandyopadhyay, S., Devi, S.L., Bhattacharyya, P., Eds.; NLP Association of India (NLPAI): Silchar, India, 2021; pp. 537–543.
55. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.E.; Lomeli, M.; Hosseini, L.; Jégou, H. The FAISS Library. *arXiv* **2024**, arXiv:2401.08281.
56. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. LLAMa-2: Open Foundation and Fine-Tuned Chat Models. 2023. Available online: https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/ (accessed on 31 December 2024).
57. Al-Dujaili Al-Khazraji, M.J.; Ebrahimi-Moghadam, A. An Innovative Method for Speech Signal Emotion Recognition Based on Spectral Features Using GMM and HMM Techniques. *Wirel. Pers. Commun.* **2024**, *134*, 735–753. [CrossRef]

58. Jiao, Z.; Ji, Y.; Gao, P.; Wang, S.H. Extraction and Analysis of Brain Functional Statuses for Early Mild Cognitive Impairment Using Variational Auto-Encoder. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 5439–5450. [CrossRef]
59. Al-Ansari, N.; Al-Thani, D.; Al-Mansoori, R.S. User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review. *Hum. Behav. Emerg. Technol.* **2024**, *2024*, 4628855. [CrossRef]