Unsupervised Thematic Context Discovery for Explainable AI in Fact Verification: Advancing the CARAG Framework

Manju Vallayil[1][a], Parma Nand[1][b], Wei Qi Yan[1][c] and Héctor Allende-Cid[2,3,4][d]

[1]*School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand*

[2]*Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340025, Chile*

[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*

Keywords: Explainable AI(XAI), Automated Fact Verification (AFV), Retrieval Augmented Generation (RAG), Explainable AFV, Fact Checking.

Abstract: This paper introduces CARAG-u, an unsupervised extension of the Context-Aware Retrieval Augmented Generation (CARAG) framework, designed to advance explainability in Automated Fact Verification (AFV) architectures. Unlike its predecessor, CARAG-u eliminates reliance on predefined thematic annotations and claim-evidence pair labels, by dynamically deriving thematic clusters and evidence pools from unstructured datasets. This innovation enables CARAG-u to balance local and global perspectives in evidence retrieval and explanation generation. We benchmark CARAG-u against Retrieval Augmented Generation (RAG) and compare it with CARAG, highlighting its unsupervised adaptability while maintaining a competitive performance. Evaluations on the FactVer dataset demonstrate CARAG-u's ability to generate thematically coherent and context-sensitive post-hoc explanations, advancing Explainable AI in AFV. The implementation of CARAG-u, including all dependencies, is publicly available to ensure reproducibility and support further research.

1 INTRODUCTION

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

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

[a] https://orcid.org/0000-0003-1962-8837
[b] https://orcid.org/0000-0001-6853-017X
[c] https://orcid.org/0000-0002-7443-3285
[d] https://orcid.org/0000-0003-3047-8817

Vallayil, M., Nand, P., Yan, W. Q. and Allende-Cid, H.
Unsupervised Thematic Context Discovery for Explainable AI in Fact Verification: Advancing the CARAG Framework.
DOI: 10.5220/0013683400004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 67-77

empirically proven to improve the thematic alignment of the claim with the generated post-hoc explanations. By doing so, CARAG interprets and enriches evidence selection, thereby enhancing post-hoc explanations and contributing to advancements in addressing critical XAI challenges.

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

In this paper, we introduce CARAG-u, an enhanced framework that eliminates reliance on structured annotations by dynamically deriving thematic clusters and evidence pools in an unsupervised manner. This advancement broadens CARAG-u's applicability to unstructured datasets, enabling seamless operation without predefined labels, extending its usability to open-domain settings. To evaluate its effectiveness, we benchmark CARAG-u against RAG, while acknowledging CARAG as an enhancement of RAG. Despite operating independently of pre-annotated labels, CARAG-u surpasses the RAG baseline and demonstrates competitive performance with CARAG as shown in Tables (1a) & (1b). Crucially, CARAG-u achieves this advancement while preserving CARAG's core explainability features, thereby addressing a key challenge in scaling XAI solutions for AFV systems.

For evaluation, we use the same FactVer_v2.0 dataset as employed in CARAG, available on HuggingFace[1]. Although FactVer includes theme and claim-evidence annotations, these annotations are not utilized during evidence retrieval or explanation generation in CARAG-u. Instead, they are used solely to evaluate performance, particularly to assess the thematic alignment of the generated explanations, ensuring a consistent baseline and fair comparison with CARAG. This approach isolates the impact of transitioning from a supervised to an unsupervised framework while leveraging a dataset with known properties to assess CARAG-u's scalability, thematic discovery capabilities, and relevance in evidence-based reasoning tasks. Building on these design considerations, the CARAG-u framework is designed with scalability, ensuring adaptability to advancements in XAI, RAG, and LLMs. Its modular architecture allows seamless integration of state-of-the-art techniques from LLM research and RAG innovations with minimal adaptation, keeping the framework at the forefront of explainability in AFV. The complete CARAG-u implementation, is publicly available on GitHub[2].

It is equally important to highlight that both CARAG and CARAG-u enhance transparency in AFV by addressing the critical challenge of integrating local and global XAI concepts. In the context of AFV, local explainability focuses on clarifying individual predictions, whereas global explainability encompasses diverse approaches to understanding the model's overall reasoning behavior, thereby offering a more holistic view of its decision-making logic. By integrating these perspectives, CARAG and CARAG-u provide deeper insights into how individual claims relate to the broader context of a knowledge base, where context plays a pivotal role in interpreting individual facts. However, prominent literature reviews and surveys in the intersection of XAI and AFV (Vallayil et al., 2023; Kotonya and Toni, 2020a) highlight persistent gaps in this field, especially the limited focus on achieving global transparency. Existing XAI approaches in AFV, such as transformer-based summarization (Atanasova et al., 2020; Kotonya and Toni, 2020b), logic-based models (Chen et al., 2022; Krishna et al., 2022), attention mechanisms (Popat et al., 2018; Shu et al., 2019; Amjad et al., 2023), counterfactual explanations (Dai et al., 2022; Xu et al., 2023), and methods leveraging RAG for dynamic evidence retrieval and reasoning (Wang and Shu, 2023; Singhal et al., 2024), predominantly focus on local explainability, leaving the broader challenges of achieving global transparency in AFV systems largely unaddressed. Recent surveys on LLM-based fact checking (Vykopal et al., 2024) highlight the potential of LLMs to support fact-checkers with advanced reasoning capabilities, but they do not directly address the challenge of achieving global transparency in AFV systems. To the best of our knowledge, the existing literature highlights the CARAG framework (Vallayil et al., 2025), along with its precursor work on graph-based thematic clustering for explainability in AFV (Vallayil et al., 2024), as the only prior efforts in related work explicitly addressing the integration of global explainability in AFV. These works uniquely combine a claim's local con-

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

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

text with its position within the dataset's global context. While not directly focused on global explainability, methodological parallels in the literature can be drawn to broader XAI research, such as surrogate models like LIME (Ribeiro et al., 2016), which approximate complex AI model behaviors locally using Machine Learning (ML) models to provide human-understandable instance-level explanations. In a similar vein, CARAG leverages interpretable ML-based methods to illuminate the decisions of complex AI models, bridging the gap between advanced model performance and the need for interpretability in AFV systems.

The remainder of this paper is organized as follows. The methodology section details the CARAG-u methodology, including dynamic thematic context discovery, query embedding construction, and a side-by-side depiction of CARAG and CARAG-u in Figure 1. The Experiments and Results section presents the experimental setup, with comparative evaluation against RAG and CARAG, and highlights CARAG-u's evidence-based reasoning performance through a case study. Finally, the Discussion section summarizes our findings and outlines directions for future work.

2 METHODOLOGY

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

2.1 Dynamic Thematic Context Discovery

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

Clustering is performed using a Gaussian Mixture Model (GMM) optimized via the Expectation-Maximization (EM) algorithm, as shown in Equation 1. GMM-EM was chosen for its capacity to model the dataset as a mixture of latent thematic patterns, where each pattern is represented as a Gaussian component characterized by its mean and variance (Al-Dujaili Al-Khazraji and Ebrahimi-Moghadam, 2024; Barai et al., 2022; Jiao et al., 2023; Moondra and Chahal, 2023). This probabilistic formulation enables soft clustering, which is suitable for cap-

Algorithm 1: Evidence Retrieval with CARAG-u.

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

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

Step 1: Data Preparation and Clustering
Load dataset $\mathcal{D}$
Initialize empty list all_texts and append all claims and evidences in $\mathcal{D}$
for each element $e \in all\_texts$ **do**
    Generate embedding $\text{emb}(e)$
end for
Apply GMM-EM to cluster $\text{emb}(all\_texts)$: $L = \text{GMM-EM}(\{\text{emb}(e_i)\}_{i=1}^{n}, t) \rightarrow \{C_i\}_{i=1}^{t}$
Assign cluster labels $L = \{l_1, l_2, \ldots, l_n\}$ to all texts in all_texts, where $l_i \in \{c_i, c_j, \ldots, c_t\}$ and t is the number of clusters.
Step 2: SOI Generation for Selected Claim
Determine the cluster C_k containing c_s
Initialize SOI dictionary
Extract all evidences in C_k to temporary list 'cluster_evidences' (evidence pool)
for each evidence e_k in cluster_evidences **do**
    Compute similarity:

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

    if sim $> \delta$ **then**
        Add e_k to the list for key 'refined_cluster_evidences' in SOI
    end if
end for
Step 3: Thematic Embedding Generation
Compute thematic embedding:

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

Step 4: Evidence Retrieval
Compute the Combined Embedding:

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

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

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

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

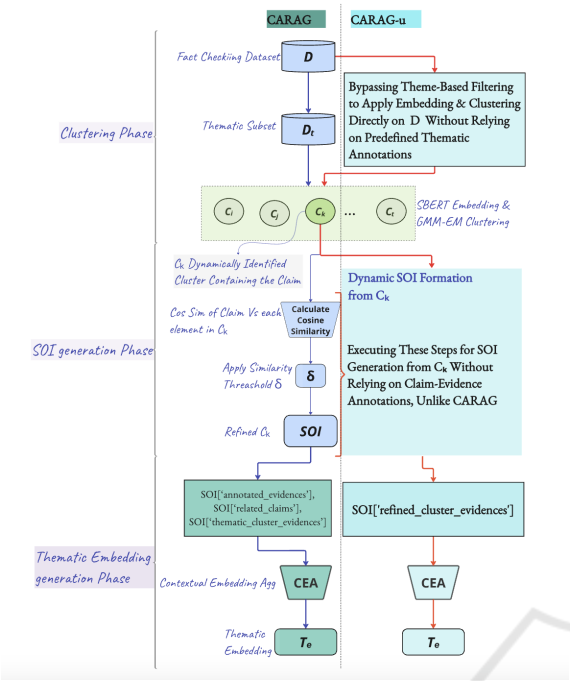


Figure 1: Comparison of CARAG and CARAG-u across clustering, SOI generation, and thematic embedding phases. CARAG relies on predefined thematic subsets and annotated evidence, while CARAG-u dynamically clusters the entire dataset ($\mathcal{D}$), forming SOIs and thematic embeddings without relying on either thematic annotations or claim-evidence annotations.

turing overlapping and ambiguous themes in natural language. It dynamically identifies a set of t clusters ($\{C_i\}_{i=1}^{t}$) based on inherent semantic relationships. The parameter t, representing the number of clusters, is configurable depending on the dataset and desired thematic resolution.

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

where L represents the cluster labels assigned to each embedding, and $\text{emb}(e_i)$ denotes the embedding of the i-th textual input from dataset $\mathcal{D}$. This step eliminates the reliance on predefined thematic filtering $\mathcal{T}$, used in CARAG to create $\mathcal{D}_T$ prior to clustering, marking an initial progression towards an unsupervised approach, as illustrated in the *Clustering Phase* of Figure 1. To ensure consistency and enable a fair comparison with CARAG during evaluation, the same SBERT model (sentence-transformers/all-mpnet-base-v2) and clustering algorithm (sklearn.mixture.GaussianMixture) were selected, leveraging their established effectiveness in semantic representation (Reimers and Gurevych, 2019) and clustering (Binti Kasim et al., 2021), respectively.

Subsequently, the cluster C_k containing the embedding of the selected claim c_s is identified from these clusters. A Subset of Interest (SOI), is then constructed from C_k by selecting evidences e_k that meet a similarity threshold δ, as defined in Equation 2, alongside c_s itself.

$$\mathcal{S}(C_k) = \{c_s\} \cup \{e_k \mid \text{sim}(e_k, c_s) > \delta \text{ and } e_k \in C_k\} \quad (2)$$

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

This approach differs fundamentally from CARAG, which relies on annotated evidences explicitly tied to c_s, as well as related claims c_r within the cluster and their corresponding annotated evidences referred to as "thematic_cluster_evidence", as illustrated in the *Soi generation Phase* of Figure 1. In contrast, CARAG-u considers all evidences within the cluster C_k as the evidence pool, irrespective of their claim annotations, initially referred to as "cluster_evidence". This pool serves as the basis for SOI formation, which is further refined into "refined_cluster_evidence" by applying the similarity threshold δ. By eliminating the dependency on claim-evidence annotations, CARAG-u enables an unsupervised and scalable process. This dynamic SOI formation allows CARAG-u to generalize the CARAG framework, operating effectively on datasets without thematic labels or predefined structures. By leveraging unsupervised processes, CARAG-u enhances its applicability and adaptability for thematic discovery.

2.2 Evidence Retrieval Query Construction from Discovered Contexts

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

CARAG-u then proceeds to construct a combined embedding E_{combined}, integrating thematic insights from T_e with claim-specific focus from E_{claim} (also denoted interchangeably as $emb(c_s)$) to form the retrieval query, as defined in Equation 3.

$$E_{combined} = \alpha \cdot E_{claim} + (1-\alpha) \cdot T_e,$$
$$T_e = \frac{1}{m} \sum_{j=1}^{m} emb(e_j),$$
$$e_j \in \mathcal{S}(C_k)['refined\_cluster\_evidences'] \quad (3)$$

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

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

2.3 Pipeline Summary and Implementation

Algorithm 1 outlines the methodology, detailing the steps for clustering, SOI formation, thematic embedding generation, and evidence retrieval using the combined embedding as the query. The practical application of this algorithm is demonstrated in the case study detailed in the Experiments and results section, showcasing its adaptability to varying parameter configurations. Additionally, the case study highlights the post-hoc explanations generated from the retrieved evidences, underscoring CARAG-u's effectiveness in unsupervised thematic discovery.

Figure 1 summarizes the distinctions between CARAG and CARAG-u across the clustering, SOI generation, and thematic embedding phases. While CARAG operates within predefined thematic subsets, CARAG-u bypasses theme-based filtering and directly applies clustering to the entire dataset ($\mathcal{D}$), enabling dynamic cluster formation without relying on thematic annotations. In the SOI generation phase, CARAG incorporates annotated evidences, related claims, and thematic cluster evidences, whereas CARAG-u solely relies on refined cluster evidences derived from the unsupervised clustering process. An experimental evaluation of our methodology is presented in the following section.

3 EXPERIMENTS & RESULTS

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

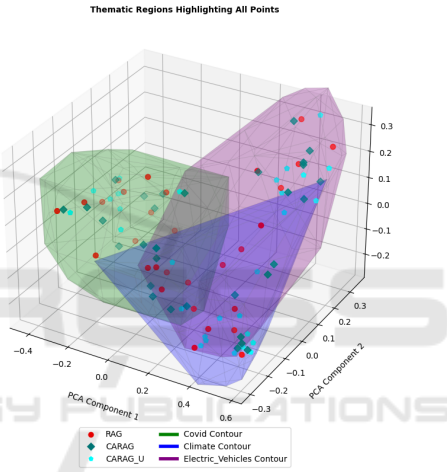


Figure 2: Visualization of explanation embeddings for RAG, CARAG, and CARAG-u within thematic boundaries (COVID: green, Climate: blue, Electric Vehicles: purple).

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

Building on the methodological framework detailed, we evaluate the effectiveness of CARAG-u in generating contextually relevant explanations. For both CARAG and CARAG-u, we first generated thematic embeddings T_e based on the respective SOIs. Using these T_e, we constructed combined embeddings E_{combined} as retrieval queries and subsequently retrieved n_{docs} evidences for a selected claim. Post-hoc explanations were then generated using these evidence sets, formatted as part of the LLM prompt[3],

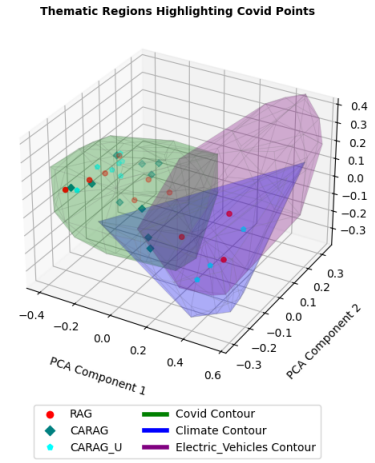
[3] prompt: <claim> + <n_{docs} evidence documents> + <instruction specifying the evidence-based claim verification and post-hoc explanation generation tasks>

with the *Llama-2-7b-chat-hf* (Touvron et al., 2023) operating in a zero-shot paradigm.

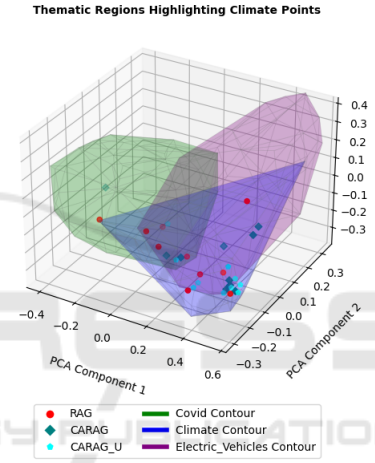
As part of this evaluation, we selected 10 claims each from three themes, COVID, Climate, and Electric Vehicles. For each claim, evidence documents were first dynamically retrieved from the FactVer dataset using RAG, and then using E_{combined}, computed with T_e corresponding to the CARAG and CARAG-u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

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

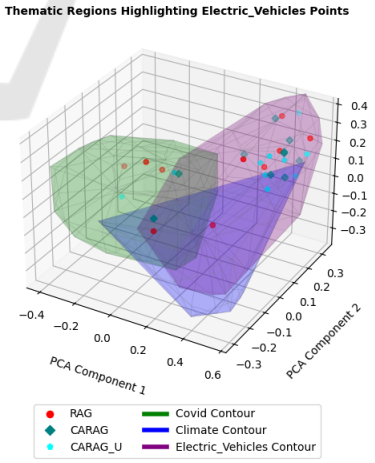
Figure 2 provides an overall visualization of the embeddings of the post-hoc explanations generated across the three frameworks. Thematic boundaries (COVID: green, Climate: blue, Electric Vehicles: purple) are depicted using 3D PCA-based convex hulls, delineating the thematic regions. RAG explanations (red circles) are broadly scattered, reflecting the generalized nature of its retrieval strategy. CARAG explanations (teal diamonds) exhibit tighter clustering within thematic boundaries due to its reliance on supervised thematic annotations. Notably, CARAG-u explanations (cyan pentagons) demonstrate comparable alignment within thematic regions, despite operating in an unsupervised manner. This demonstrates CARAG-u's ability to dynamically infer the-



(a) COVID



(b) Climate



(c) Electric Vehicles

Figure 3: Explanation embeddings disaggregated by theme (COVID, Climate, Electric Vehicles). Each subplot illustrates the alignment of RAG, CARAG, and CARAG-u explanations within thematic boundaries.

Table 1: Alignment of explanation embeddings with thematic centroids across PCA and t-SNE spaces.

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

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

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

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

matic context, highlighting its flexibility in generating contextually relevant explanations without annotated evidence. Figure 3 further disaggregates the analysis by theme, with each subplot showing explanation embeddings for one theme plotted over all thematic boundaries. For example, Figure 3a shows that the explanations generated for COVID claims predominantly fall within the green contour. CARAG explanations remain tightly clustered, reflecting their reliance on thematic supervision. Similar patterns are observed for Climate (Figure 3b) and Electric Vehicles (Figure 3c), further underscoring CARAG-u's ability to approximate the performance of CARAG while operating without evidence and thematic annotations.

Tables (1a) and (1b) complement these visualizations with quantitative insights. Table 1a reports Euclidean distances between explanation embeddings (RAG, CARAG, and CARAG-u) and their respective thematic centroids across PCA and t-SNE spaces for each theme, including average values. These distances measure how closely the embeddings align

Table 2: Case study example of post-hoc explanations generated by CARAG-u for the claim "Climate change is caused by atmospheric concentration of greenhouse gases." These explanations are based on retrieved evidence for different α values (Equation 3), with *delta* set to 0.75 (Equation 2), and t set to 10 (Equation 1). The variations in α demonstrate its impact on thematic alignment and evidence retrieval.

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

with their thematic regions, with lower values indicating better alignment. CARAG-u consistently achieves tighter alignment than RAG across most themes and embedding spaces. Notable improvements are observed for Climate, where CARAG-u achieves the lowest t-SNE distance (14.8924), while demonstrating competitive performance with CARAG for COVID and Electric Vehicles. The inclusion of t-SNE distances highlights CARAG-u's ability to capture local relationships, enhancing its contextual alignment. Table (1b) shows differences in alignment (CARAG minus RAG & CARAG-u minus RAG distance in PCA and in t-SNE spaces). Negative values indicate superior alignment over RAG. CARAG-u exhibits substantial improvements in Climate (-0.082 in PCA, -0.1788 in t-SNE) and Electric Vehicles (-0.0946

in PCA, -0.1603 in t-SNE). While CARAG benefits from supervised thematic embeddings, CARAG-u performs competitively, as evidenced by overall averages (-0.0452 for PCA and -0.1302 for t-SNE). In summary, CARAG-u balances unsupervised adaptability with thematic alignment, offering an alternative to RAG in contexts where structured annotations are unavailable. These results validate its robustness and extend the explainability framework established by CARAG to broader, less structured domains, advancing the goal of unsupervised fact verification.

3.2 Case Study

Building on the broader framework comparison, we conducted a focused case study to assess CARAG-u's performance in dynamically generating thematic embeddings and their influence on evidence retrieval and explanation generation. This evaluation, centered on varying the weighting parameter α (Equation 3), used the claim, "Climate change is caused by atmospheric concentration of greenhouse gases" (Claim ID: 44)[4] from the FactVer dataset.

The case study followed the methodological steps outlined in Algorithm 1, with the following parameter configuration: $\mathcal{D} = FactVer$, $c_s = 44$, $\delta = 0.75$, $n_{\text{docs}} = 6$, $\alpha \in \{1, 0.75, 0.50, 0.25\}$, and $t = 10$. Anchoring the evaluation to this structured process highlights CARAG-u's adaptability to varying parameter configurations, where α determines the extent to which the thematic context T_e is incorporated into the evidence retrieval query E_{combined} as defined in Equation 3. Consequently, when $\alpha = 1$, CARAG-u relies solely on E_{claim} for evidence retrieval, mimicking the behavior of a traditional RAG framework as our baseline setup. Conversely as α decreases, the influence of T_e increases, incorporating broader thematic context (global context) at the expense of claim specificity.

Table 2 illustrates the evolution of post-hoc explanations generated by CARAG-u across varying α configurations. At α = 1 or at the baseline RAG setup, producing explanations disconnected from the thematic context of the claim (climate change), such as an irrelevant focus on labor market impacts. As α decreases, the thematic embedding T_e increasingly influences retrieval process, aligning the retrieved evidence and consequently the explanations, more closely with the claim. Notably, at α = 0.25, the explanations emphasize industrial de-carbonization schemes and greenhouse gas emissions, directly supporting the claim. While its predecessor CARAG also supports adaptability in evidence retrieval through varying α, CARAG-u stands out by achieving this in a fully unsupervised manner.

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

4 DISCUSSION

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

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

REFERENCES

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

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

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

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

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

Binti Kasim, F. A., Pheng, H. S., Binti Nordin, S. Z., and Haur, O. K. (2021). Gaussian Mixture Model - Expectation Maximization Algorithm for Brain Images. In *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pages 1–5.

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

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

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

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

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

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

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

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

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

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

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

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

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

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

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

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

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

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

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

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

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

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

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