KG-CRAFT: Knowledge Graph-based Contrastive Reasoning with LLMs for Enhancing Automated Fact-checking

Vítor N. Lourenço[1,2], Aline Paes[1], Tillman Weyde[3], Audrey Depeige[2], Mohnish Dubey[2]

[1]Universidade Federal Fluminense, [2]Amazon, [3]City St George's, University of London
Correspondence: vitorlourenco@id.uff.br, vitornl@amazon.co.uk

Abstract

Claim verification is a core component of automated fact-checking systems, aimed at determining the truthfulness of a statement by assessing it against reliable evidence sources such as documents or knowledge bases. This work presents KG-CRAFT, a method that improves automatic claim verification by leveraging large language models (LLMs) augmented with contrastive questions grounded in a knowledge graph. KG-CRAFT first constructs a knowledge graph from claims and associated reports, then formulates contextually relevant contrastive questions based on the knowledge graph structure. These questions guide the distillation of evidence-based reports, which are synthesised into a concise summary that is used for veracity assessment by LLMs. Extensive evaluations on two real-world datasets (LIAR-RAW and RAWFC) demonstrate that our method achieves a new state-of-the-art in predictive performance. Comprehensive analyses validate in detail the effectiveness of our knowledge graph-based contrastive reasoning approach in improving LLMs' fact-checking capabilities.

1 Introduction

The digital transformation has reshaped how society consumes and shares information, posing new challenges to information integrity (Haider and Sundin, 2022). *The Reuters Institute Digital News Report 2024* highlights how social media has fragmented the news ecosystem (Newman et al., 2024). Despite expanding access and engagement, this shift has also fuelled the spread of misinformation (Valenzuela et al., 2019).

Misinformation is particularly concerning in high-stakes contexts such as elections and public health crises, where it can cause serious societal harm. Consequently, the demand for more effective and scalable fact-checking methods has driven the emergence of Automated Fact-Checking (AFC) systems (Zhou and Zafarani, 2020; Alam et al., 2022; Guo et al., 2021; Eldifrawi et al., 2024).

AFC is designed to assess the veracity of claims by retrieving, analysing, and reasoning over relevant evidence from reliable sources (Wu et al., 2025). Early approaches relied on classification and evidence retrieval pipelines (Shu et al., 2019; Kotonya and Toni, 2020; Atanasova et al., 2020). Whilst the integration of relational structures and knowledge bases (Lourenço and Paes, 2022; Whitehouse et al., 2022; Lourenço et al., 2025; Huang et al., 2025; Chen et al., 2025), and hierarchical architectures (Yang et al., 2022) improved performance, these methods lacked the scalability and adaptability that LLMs later demonstrated. Recent LLM-based approaches have achieved significant advances in AFC through the integration of external knowledge (Cheung and Lam, 2023; Guo et al., 2023) and the introduction of retrieval mechanisms (Singal et al., 2024; Zhang and Gao, 2024; Yue et al., 2024). However, these solutions often lack structured reasoning mechanisms (Liu et al., 2024), which can lead to unreliable verification processes. Contrastive reasoning (Jacovi et al., 2021; Paranjape et al., 2021) has demonstrated effectiveness in enhancing model interpretability and decision-making, yet its application to fact-checking remains underexplored.

To address the aforementioned open challenges and ultimately enhance AFC capabilities, we propose a method for improving *claim verification* within the AFC pipeline. We focus on claim verification in a bounded context, where each claim is accompanied by a predefined set of associated reports. This setting reflects numerous domains, *e.g.*, legal document review (Zheng et al., 2021), financial auditing (Zhu et al., 2021), and scientific peer review (Wadden et al., 2020), where analysis is restricted to a specific corpus. Consequently, the

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6419–6439

core challenge shifts from open-domain evidence retrieval to reasoning effectively over the available information to determine a claim's veracity. Specifically, we propose *Knowledge Graph-based Contrastive Reasoning for Automated FacT verification* (KG-CRAFT).

Motivated by findings in both cognitive science and natural language processing (Schuster et al., 2021; Buçinca et al., 2025), we leverage contrastive reasoning for our automated fact-checking task. Prior work highlights that verification requires distinguishing whether a claim is supported or contradicted (Thorne et al., 2018; Aly et al., 2021), and that contrastive explanations align more closely with human reasoning (Miller, 2019). Moreover, contrastive learning methods have proven effective in enforcing meaningful semantic distinctions (Chen et al., 2020; Gao et al., 2021). We therefore introduce contrastive questions into the fact-checking pipeline, encouraging models not only to assess whether evidence supports a claim but also to explicitly consider alternatives, thereby promoting more robust verification.

However, generating meaningful contrasts from unstructured text alone is a non-trivial challenge. Without a structured representation of the underlying facts and their relationships, contrastive questions may focus on superficial linguistic differences rather than semantically significant distinctions, leading to less informative or even arbitrary contrasts (Bhattacharjee et al., 2022). By explicitly encoding entities and their semantic relations, a knowledge graph (KG) provides a structured means of identifying candidates for contrast (Liu et al., 2019). This structure guides the model to formulate questions that explore meaningful conceptual distinctions (*e.g.*, contrasting one entity with another of the same type), ensuring that the reasoning process is both more robust and more aligned with human-like. This approach helps to ensure that contrasts are grounded in verification-relevant relationships rather than being purely text-driven.

To implement this structured contrastive approach, KG-CRAFT first decomposes each claim and its associated reports into entities and their relationships to construct the KG. After, it formulates and selects contrastive questions based on the KG structure, aiming to maximise both diversity and contextual relevance to the claim. The set of questions is then answered using the input reports, generating a new, contextually relevant, evidence-based information set. Inspired by the scalability and demonstrated performance of LLMs in automated fact-checking (Cheung and Lam, 2023; Wang et al., 2024a; Xiong et al., 2025), this new set is consolidated into a concise summary representing a distilled version of the input reports, which is then used to assess the veracity of the claim.

The main contributions of this work are: *(i)* a novel Knowledge Graph-based Contrastive Reasoning method that enhances LLM capabilities in AFC; *(ii)* state-of-the-art performance on two real-world fact-checking datasets; and *(iii)* a comprehensive ablation study analysing the proposed components for AFC.

2 Related Work

This section situates our contribution within prior work on AFC, and LLM-based claim verification. Foundational concepts are deferred to Appendix A, which provides formal definitions and notation for (i) *contrastive explanations and reasoning*, and (ii) *knowledge graph construction with LLMs*; these concepts are developed in greater detail in the appendix. Appendix A also presents an extended review of related work, including further discussion of KG-based approaches to fact-checking.

Non-generative Automated Fact-checking Classical AFC models encode claims and documents using text embeddings, and verify them via supervised classifiers. Notable systems include dEFEND, which employs sentence–comment co-attention for news and user comments (Shu et al., 2019); SBERT-FC, which introduced the PubHealth dataset, and an explainability analysis (Kotonya and Toni, 2020); and GenFE/GenFE-MT, which jointly optimise veracity prediction and explanation generation (Atanasova et al., 2020). CofCED proposes a hierarchical encoder with cascaded evidence selectors for multi-source reports (Yang et al., 2022). Incorporating KGs into pretrained models (*e.g.*, via Wikidata) improves accuracy, especially for political claims (Whitehouse et al., 2022); a recent survey comprehensively reviews KG-based AFC (Qudus et al., 2025).

Fact-checking Using LLMs LLMs have become central to AFC, yet their reliability remains constrained by training coverage and hallucinations (Wang et al., 2024b; Augenstein et al., 2024). Early systems augment LLMs with structure or external evidence: FactLLaMA couples instruction-following with retrieval (Cheung and Lam, 2023); IKA builds example graphs for verification and explanation (Guo et al., 2023); TELLER integrates human expertise with LLM reasoning (Liu et al., 2024); defence-style frameworks partition evidence into competing narratives for robust verification (Wang et al., 2024a); and CorXFact models claim–evidence correlations (Tan et al., 2025). Retrieval-augmented generation (RAG), ranging from basic RAG pipelines (Singal et al., 2024), to retrieval optimised with fine-grained feedback (Zhang and Gao, 2024), and architectures targeting evidence retrieval plus contrastive argument synthesis (Yue et al., 2024), has become an increasingly prominent approach. Other recent directions include iterative verification for scalability (FIRE) (Xie et al., 2025), and the handling of zero-day manipulations via real-time context retrieval (Meng et al., 2025).

3 Knowledge Graph-based Contrastive Reasoning

We introduce a novel approach that leverages knowledge graphs to fuel contrastive reasoning and enhance LLM fact-checking capabilities: *Knowledge Graph-based Contrastive Reasoning for Automatic FacT Verification* (KG-CRAFT). Our work focuses on the claim verification component of automated fact-checking, integrating contrastive reasoning into the verification process through the use of structured evidence to generate contextually relevant contrastive queries, thereby guiding more accurate claim classification. We begin with the claim verification task formulation.

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

Next, we present the components of KG-CRAFT, depicted in Figure 1: knowledge graph construction from the textual input (the claim and its associated reports) (Section 3.1); contrastive reasoning, comprising contrastive question generation, answer generation, and prompt-based answer summarisation (Section 3.2); and claim veracity verification (Section 3.3).

3.1 Knowledge Graph Extraction

The first phase of KG-CRAFT extracts entities and relationships from $\mathcal{C}$ and $\mathcal{R}_\mathcal{C}$ and uses them to construct a knowledge representation of the input. We draw inspiration from prior work (Zhu et al., 2024; Zhang and Soh, 2024) and leverage LLMs for knowledge graph construction. Through phased few-shot prompting, we instrument the LLM to first identify entities $\mathcal{E}$, then label them according to their conceptual categories $\mathbb{C}$, to give them semantic meaning and enable disambiguation. Next, we identify the relationships $\mathcal{R}$ that relate the identified entities (prompt details in Appendix D.1), forming a set of triples $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. The unified resulting sets constitute the input knowledge graph $\mathcal{G}_{\mathcal{C},\mathcal{R}_\mathcal{C}} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathbb{C})$.

3.2 Contrastive Reasoning

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

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

To ensure that the formulated questions are individually relevant and collectively diverse, we adopt

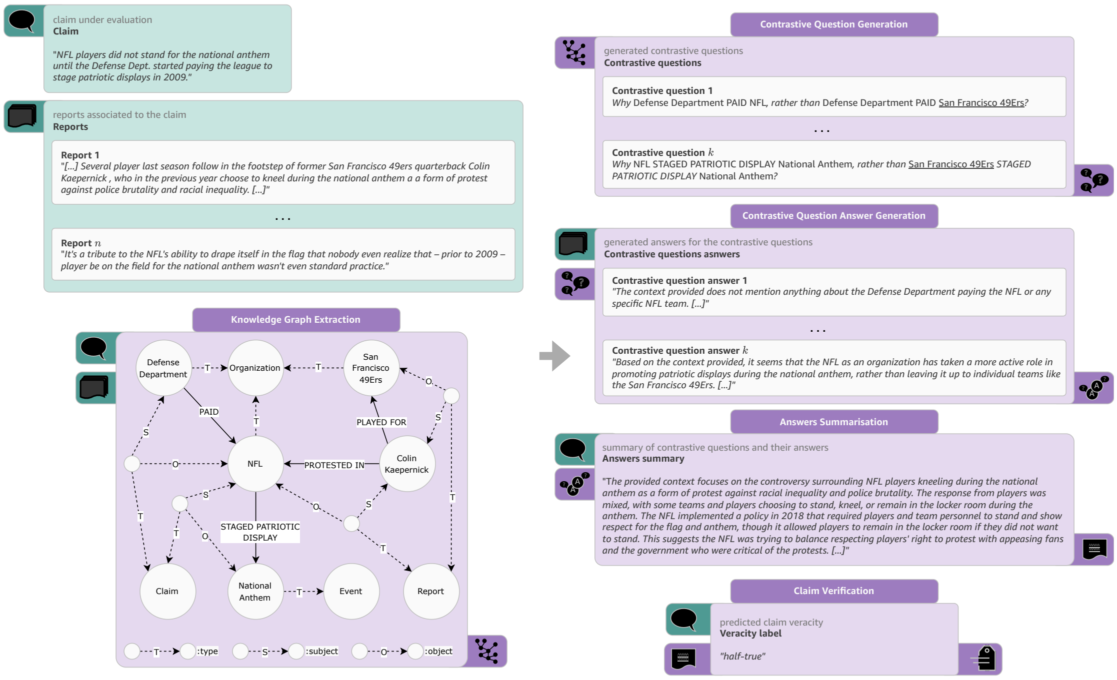

Figure 1: Overview of the KG-CRAFT framework for automated fact-checking. The method comprises three main phases: (1) knowledge graph extraction from the claim and associated reports, (2) contrastive reasoning, including contrastive question formulation, answer generation, and answer summarisation, and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

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

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

$$\mathsf{q}_i = \underset{\substack{\mathsf{q} \in \mathcal{Q}_{\text{Em}} \\ \setminus \mathcal{Q}_{ranked}}}{\arg\max} \left[\text{Sim}(\mathsf{q}, \mathsf{q}_\theta) - \max_{\mathsf{q}' \in \mathcal{Q}_{ranked}} \text{Sim}(\mathsf{q}, \mathsf{q}') \right], \tag{1}$$

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

Contrastive Question Answer Generation The second step of the process answers the previously formulated contrastive questions $\mathcal{Q}^{\mathtt{K}}_{ranked}$ through a structured query-response prompt mechanism p_{ag} (available in Appendix D.2). An LLM analyses and generates information from the claim-associated reports – prompting it to answer independently the contrastive questions based on the set of reports $\mathcal{R}_{\mathcal{C}}$ associated with the claim. This process generates a set of answers $\tilde{\mathcal{A}} = \{LLM_{p_{\text{ag}}}(q, \mathcal{R}_{\mathcal{C}}) \mid q \in \mathcal{Q}^{\mathtt{K}}_{ranked}\}$, where each answer is directly derived from the reports, maintaining traceability between claim, reports, contrastive questions, and answers. Our goal by highlighting the contrastive elements in the generated answers is to highlight the key evidence supporting the claim's veracity during the reasoning process.

Answers Summarisation The final step of the Contrastive Reasoning process involves aggregating the question-answer pairs into a concise, evidence-based summary. From the claim $\mathcal{C}$ and the paired contrastive questions and answers from $\mathcal{Q}^{\mathtt{K}}_{ranked}$ and $\tilde{\mathcal{A}}$, respectively, we prompt p_{as} (Appendix D.3) to an LLM to generate a concise paragraph that relates all contrastive question-answer pairs. This summarisation step, represented as $A_{\mathcal{C}} = LLM_{p_{as}}(\mathcal{C}, \{(q_i, a_i) \mid q_i \in \mathcal{Q}^{\mathtt{K}}_{ranked}, a_i \in \tilde{\mathcal{A}}\})$, ensures that key contrasting elements and supporting evidence are presented in a structured summary, producing a distilled source of informa-

Algorithm 1 Define Contrastive Questions

Require:
$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathbb{C})$ ▷ Claim and reports KG; entities $\mathcal{E}$, relations $\mathcal{R}$, triples $\mathcal{T}$, and entities' classes $\mathbb{C}$
$\mathcal{T}_{\text{claim}}$ ▷ Claim's extracted triples; $\mathcal{T}_{\text{claim}} \subseteq \mathcal{T}$
$\mathtt{K}$ ▷ Maximum # of Contrastive Questions
Ensure:
$\mathcal{Q}^{\mathtt{K}}_{ranked}$ ▷ Set of top $\mathtt{K}$ most relevant and diverse contrastive questions for a claim and its reports

$\mathcal{Q} \leftarrow \emptyset$
for each $\mathtt{t} = \{h, r, t\} \in \mathcal{T}_{\text{claim}}$ **do**
    $h_c \leftarrow \tau(h)$ ▷ where $\tau : \mathcal{E} \rightarrow \mathbb{C}$
    $t_c \leftarrow \tau(t)$
    $H_{contr} \leftarrow \{h \mid h \in \mathcal{E}, \tau(h) = h_c\} \setminus \{h\}$
        ▷ Set of entities of the same class as *head*
    $T_{contr} \leftarrow \{t \mid t \in \mathcal{E}, \tau(t) = t_c\} \setminus \{t\}$
        ▷ Set of entities of the same class as *tail*
    for each $h' \in H_{contr}$, $t' \in T_{contr}$ **do**
        $q_h \leftarrow \textsc{FormulateQuestion}(\mathtt{t}, h')$
        $q_t \leftarrow \textsc{FormulateQuestion}(\mathtt{t}, t')$
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q_h, q_t\}$
    end for
end for
$\mathcal{Q}_{ranked} \leftarrow \textsc{ReRank}(\mathcal{Q})$ ▷ Rank contrastive questions based on Equation (1)
$\mathcal{Q}^{\mathtt{K}}_{ranked} \leftarrow (\mathcal{Q}_{ranked})_{1:K}$ ▷ Select first K elements from ranked set

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

tion. The resulting summary $A_{\mathcal{C}}$ emphasises key contrasting facts whilst abstracting non-essential information, preserving semantic links between critical evidence and the claim to create a focused source of verification.

3.3 Verification of Claim Veracity

The last phase of KG-CRAFT assesses the claim's veracity $\mathcal{V}_{\mathcal{C}}$. For that, we prompt p_{cv} (see Appendix D.4) an LLM, containing the original claim $\mathcal{C}$ and the produced summary $A_{\mathcal{C}}$ as the only source of evidence, mitigating potential noise in the original reports $\mathcal{R}_{\mathcal{C}}$. The prompt p_{cv} also includes the possible labels and their descriptions. In this way, the LLM acts as a classifier to infer the claim's veracity, represented as $\mathcal{V}_{\mathcal{C}} = LLM_{p_{\text{cv}}}(\mathcal{C}, A_{\mathcal{C}})$, ensuring that the veracity assessment is based on the distilled evidence produced through our Contrastive Reasoning method.

4 Experiments and Results

This section is guided by the following research questions: **RQ1** How effective is KG-CRAFT in claim verification compared to state-of-the-art methods? **RQ2** How beneficial are KG-based contrastive questions compared to purely LLM-generated contrastive questions? **RQ3** What is the effect of the number of contrastive questions $\mathtt{K}$ on claim verification with KG-CRAFT? **RQ4** How effective is KG-CRAFT with Small Language Models (SLMs) compared to LLMs? To answer these questions, we evaluate KG-CRAFT on two real-world fact-checking benchmarks and compare the results with baseline methods, and perform several ablation studies. Additional experiments and ablation studies can be found in Appendix C.

4.1 Experimental Settings

Datasets We evaluated the proposed approach on two publicly available fact-checking datasets: LIAR-RAW and RAWFC, (refer Appendix B.2 for datasets statistics). LIAR-RAW (Yang et al., 2022) extends LIAR-PLUS (Alhindi et al., 2018) and contains fine-grained claims from Politifact[1] with six veracity classes (`pants-fire`, `false`, `barely-true`, `half-true`, `mostly-true`, `true`) along with their relevant reports. RAWFC (Yang et al., 2022) contains claims collected from Snopes[2] on various topics with three veracity classes (`false`, `half`, `true`) and their associated reports retrieved using claim keywords.

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

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

Llama 2 7B and ChatGPT 3.5 Turbo.

Implementation Details We extracted the KGs (Section 3.1) of both datasets utilising Claude 3 Haiku. Further, KG-CRAFT is instantiated and evaluated (Section 4.2) using Claude 3.5 Sonnet (KG-CRAFT$_{C3.5}$), Claude 3.7 Sonnet (KG-CRAFT$_{C3.7}$), and Llama 3.3 70B (KG-CRAFT$_{L3.3}$) (refer to Appendix B for more implementation details).

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

4.2 Claim Verification Evaluation

We first address **RQ1** by evaluating the performance of KG-CRAFT in verifying claim veracity. As shown in Table 1, our method – in the instances KG-CRAFT$_{C3.7}$ and KG-CRAFT$_{L3.3}$ (all using five contrastive questions $\mathtt{K} = 5$) – consistently outperforms all other methods on both datasets. KG-CRAFT$_{C3.5}$ outperforms all comparator methods except for DelphiAgent$_{\text{gpt-4o}}$) on the RAWFC dataset. KG-CRAFT$_{L3.3}$ improves the F1-score by 44 percentage points (pp) on the LIAR-RAW dataset and 13 pp on the RAWFC dataset, compared to the second best performing methods, L-Defense (Wang et al., 2024a) and (DelphiAgent (Xiong et al., 2025)), respectively. Compared to their Naïve LLM counterparts, KG-CRAFT$_{C3.5}$, KG-CRAFT$_{C3.7}$, and KG-CRAFT$_{L3.3}$, show F1 performance gains of 32, 44, and 42 pp on LIAR-RAW and 11, 12, and 27 pp on RAWFC.

4.3 Ablation Studies

To evaluate KG-CRAFT's components and reveal its strengths and limitations, we conduct three ablation studies. We first compare KG-based and LLM-generated contrastive questions to evaluate the benefits of structured question formulation (**RQ2**). Then, we analyse the effect of varying the number $\mathtt{K}$ of contrastive questions (**RQ3**). Finally, we assess whether KG-CRAFT boosts the performance of Small Language Models (SLMs) by comparing their results to Claude 3.7 Sonnet and KG-CRAFT$_{C3.7}$ (**RQ4**).

4.3.1 Impact of Using LLM-generated Contrastive Questions

To answer **RQ2**, we replaced the Contrastive Question Formulation component of KG-CRAFT (Section 3.2) by a few-shot prompt that, given the claim, reports, and examples, requests the LLM to generate $k = 5$ contrastive questions (prompt details in Appendix D.5). Results depicted on Table 2 show fact-checking F1-score, and macro and weighted AlignScore and RQUGE for the LIAR-RAW dataset.

AlignScore is a metric based on a general function of information alignment to perform automatic factual consistency evaluation of text pairs (Zha et al., 2023). In our evaluation, we use AlignScore (here called macro AlignScore) to measure the information alignment between the text piece generated at the answer summarisation step (Section 3.2) with the original claim. *RQUGE* is a reference-free evaluation metric for generated questions, based on the corresponding context and answer (Mohammadshahi et al., 2023). In our evaluation, we use RQUGE (here called macro RQUGE) to measure the acceptability score of a formulated contrastive question (Section 3.2), given a corresponding context (the claim) and an answer (the answer generated from the contrastive question Section 3.2). We further extend AlignScore and RQUGE (here called weighted AlignScore and weighted RQUGE) to weight their results by the distance of the predicted claim veracity class from the ground truth claim class. For that, we assigned each class an integer numerical value ranging from 1 (`pants-fire`) to 6 (`true`) and measured the mean squared error (MSE) between the predicted class. After, we weight AlignScore and RQUGE macro scores using the obtained MSE (details of the AlignScore and RQUGE weighted scores are in Appendix B.3 and Appendix B.4, respectively).

As depicted in Table 2, the contrastive questions generated by the LLMs, results in lower information alignment (AlignScore) between the answer summary $A_{\mathcal{C}}$ and claim $\mathcal{C}$ and lower question acceptability score (RQUGE) between formulated contrastive questions $Q^{\mathtt{K}}_{ranked}$, generated answer $\tilde{A}$, and claim $\mathcal{C}$. As a consequence, the lower information alignment and question acceptability score impact the fact-checking capability, resulting in a lower fact-checking F1-score.

4.3.2 Impact of the Number of Contrastive Questions

To address **RQ3**, we investigate how varying the number of contrastive questions k affects the performance of KG-CRAFT. We conduct this analysis

Table 1: Fact-checking results (%) on the RAWFC and LIAR-RAW datasets.

Method	LIAR-RAW			RAWFC		
	Pr	Re	F1	Pr	Re	F1
Traditional						
dEFEND (Shu et al., 2019)	23.09	18.56	17.51	44.93	43.26	44.07
SBERT-FC (Kotonya and Toni, 2020)	24.09	22.07	22.19	51.06	45.92	45.51
GenFE (Atanasova et al., 2020)	28.01	26.16	26.49	44.92	44.74	44.43
GenFE-MT (Atanasova et al., 2020)	18.55	19.90	15.15	45.64	45.27	45.08
CofCED (Yang et al., 2022)	29.48	29.55	28.93	52.99	50.99	51.07
EExpFND (Wang et al., 2025)	29.91	29.93	29.58	54.43	53.49	53.61
Naïve LLM						
Llama 2 7B (Wang et al., 2024a)	17.11	17.37	15.14	37.30	38.03	36.77
ChatGPT 3.5 Turbo (Wang et al., 2024a)	25.41	27.33	25.11	47.72	48.62	44.43
Claude 3.5 Sonnet	29.25	27.83	28.52	54.85	55.04	54.94
Claude 3.7 Sonnet	28.26	26.63	27.42	55.90	57.12	56.50
Llama 3.3 70B	47.21	23.52	31.40	53.16	55.04	54.08
Specialised LLM						
FactLLAMA (Cheung and Lam, 2023)	32.32	31.57	29.98	53.76	54.00	53.76
FactLLAMA$_{\text{know}}$ (Cheung and Lam, 2023)	32.46	32.05	30.44	56.11	55.50	55.65
L-Defense$_{\text{LLAMA2}}$ (Wang et al., 2024a)	31.63	31.71	31.40	60.95	61.00	60.12
L-Defense$_{\text{ChatGPT}}$ (Wang et al., 2024a)	30.55	32.20	30.53	61.72	61.01	61.20
DelphiAgent$_{\text{gpt-4o-mini}}$ (Xiong et al., 2025)	32.79	22.33	26.03	68.53	63.95	64.68
DelphiAgent$_{\text{gpt-4o}}$ (Xiong et al., 2025)	31.33	28.36	28.36	68.05	68.03	68.04
KG-CRAFT (ours)						
KG-CRAFT$_{C3.5}$	62.70	59.38	60.99	67.10	66.25	66.67
KG-CRAFT$_{C3.7}$	**73.92**	**<u>70.86</u>**	**72.36**	**69.37**	**68.59**	**68.98**
KG-CRAFT$_{L3.3}$	**<u>77.38</u>**	**70.67**	**<u>73.87</u>**	**<u>81.63</u>**	**<u>81.53</u>**	**<u>81.58</u>**

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

Table 2: Ablation study on the LIAR-RAW dataset comparing the quality of KG-based in contrast to LLM-based formulated contrastive questions. Metrics include Fact-checking F1-score, and macro (-M) and weighted (-W) values for AlignScore (AS) and RQUGE (RQ).

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

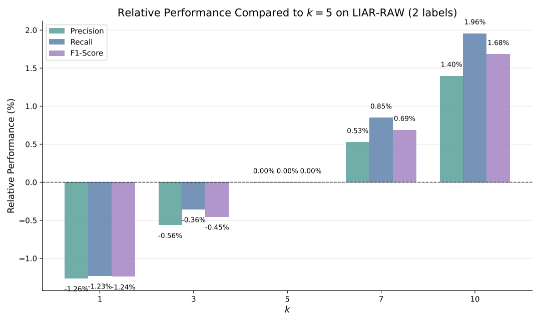

Figure 2: Impact of varying the number of contrastive questions k on fact-checking performance.

on a binary classification variant of LIAR-RAW, where the original six labels are mapped into two classes: {`pants-fire`, `false`, `barely-true`} as `false`, and {`half-true`, `mostly-true`, `true`} as `true` (details in Appendix B.2). We evaluate KG-CRAFT$_{C3.5}$ using $k \in \{1, 3, 5, 7, 10\}$ contrastive questions, with $\mathtt{K} = 5$ serving as our baseline for relative performance comparison.

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

4.3.3 Using Small Language Models

We investigate whether our KG-based contrastive reasoning methodology can enhance the performance of Small Language Models (SLMs), as raised in question **RQ4**. We evaluate four SLMs: two with less than 600M parameters (SmolLM2 135M (Allal et al., 2025) and Qwen3 0.6B (Yang et al., 2025)) and two with less than 2B parameters (SmolLM2 1.7B (Allal et al., 2025) and Qwen3 1.7B (Yang et al., 2025)). We compare their F1 performance against Naïve Claude 3.7 Sonnet and KG-CRAFT$_{C3.7}$.

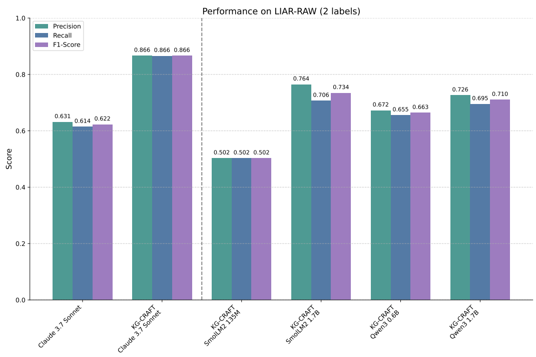

Figure 3: Performance comparison of Small Language Models incorporated within KG-CRAFT against larger models.

The results in Figure 3 show that our framework significantly enhances SLMs' fact-checking performance. Notably, SmolLM2 1.7B achieves an F1-Score of 73.40%, substantially outperforming Naïve Claude 3.7 Sonnet (62.22%). Even the smaller models show promising results, with Qwen3 0.6B reaching an F1-Score of 66.34%, which surpasses Naïve Claude 3.7 Sonnet despite its much smaller size.

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

5 Conclusion

This paper presents KG-CRAFT, a novel method for claim veracity classification via knowledge graph-based contrastive reasoning with LLMs. Our results demonstrate that KG-CRAFT outperforms state-of-the-art methods on two real-world datasets (LIAR-RAW and RAWFC). Our empirical analysis indicates that the contrastive reasoning method generates contextually relevant and evidence-based information that aids in assessing claim veracity. We investigate the effect of the number of contrastive questions used in the process, finding that more questions lead to better results, but even small numbers lead to competitive performance. Even with Small Language Models, our approach proves competitive with LLM baselines. For future work, we propose an extended evaluation with other domains and the generation of explanations for full AFC based on contrastive answer summaries.

Limitations

Whilst our results demonstrate the effectiveness of the proposed framework, some limitations remain. First, we do not qualitatively verify the intermediate components of our pipeline, such as the knowledge graph construction step, which is particularly sensitive and central to the method's overall performance. Additionally, we rely on a fixed set of LLMs for the intermediate tasks, though alternative models or fine-tuned approaches could potentially yield improved results. Our evaluation is limited to two datasets; however, these are widely adopted benchmarks in the fact-checking literature and provide a meaningful basis for comparison. We also acknowledge the reliance on expensive LLMs throughout the process, which may not be accessible to all users. However, our results also demonstrate the potential of the proposed framework to enhance the performance of smaller language models that require significantly fewer computational resources.

Although we claim to improve the results of Specialised LLMs-based AFC, we are aware of a non-negligible potential threat to the validity of our results: the use of different families of LLMs in previous and in our work. The results may therefore be influenced by variations in model capabilities, and different outcomes could emerge if their LLMs or ours were replaced. However, we were unable to modify the original systems' LLMs or employ legacy models in our experiments. This highlights a broader challenge for the community: ensuring reproducibility and fair comparison in an ecosystem where new and more capable LLMs are continuously emerging. Nonetheless, we presented ablation studies of KG-CRAFT with LLMs of limited size (SLMs), showing that it significantly enhances their performance.

Furthermore, all experiments are conducted in the English language. Although our method is designed to be language-agnostic, performance may vary across languages due to potential limitations in intermediate components such as entity linking, relation extraction, or question generation. Future work should explore broader language coverage, dataset diversity, and deeper analysis of intermediate outputs.

Ethical Considerations

Automated fact-checking with LLMs raises important ethical concerns, particularly due to their ability to produce fluent yet incorrect or fabricated outputs. When used to synthesise evidence or generate contrastive questions, such errors can mislead users or amplify misinformation. Condensing large evidence sets also involves decisions about what to include or omit; if not carefully managed, this can result in oversimplification or the exclusion of critical context, compromising factual accuracy. Moreover, whilst contrastive reasoning may improve interpretability, the overall decision-making process remains difficult to audit, with limited traceability from evidence to claim and limited justification for model outputs. Automated fact-checking systems may also be misinterpreted as authoritative, reinforcing confirmation bias. It is therefore essential to clearly communicate model limitations and maintain human oversight. We emphasise the importance of fairness-aware design, transparent evidence attribution, and rigorous human evaluation to mitigate these risks.

Use of Generative AI

During the preparation of this work, the authors used the Claude family models and the Amazon Nova family models to grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Acknowledgments

The second author thanks the support of FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, CNPq (National Council for Scientific and Technological Development), (grant #307088/2023-5) and the National Institutes of Science and Technology (INCT): IAIA (grant #406417/2022-9), TILD-IAR (grant #408490/2024-1), and IAPROBEM (grant #408589/2024-8).

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proc. 29th Int. Conf. Comput. Linguist., COLING*.

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

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

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

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Amrita Bhattacharjee, Mansooreh Karami, and Huan Liu. 2022. Text transformations in contrastive self-supervised learning: A review. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5394–5401. ijcai.org.

Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 1024:1–1024:25. ACM.

Bohan Chen and Andrea L Bertozzi. 2023. Autokg: Efficient automated knowledge graph generation for language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3117–3126. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. 2025. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*.

Islam Eldifrawi, wi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*.

Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824.

Hao Guo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2023. Interpretable fake news detection with graph evidence. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Zhijiang Guo, M. Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Jutta Haider and Olof Sundin. 2022. *Paradoxes of media and information literacy: The crisis of information*. Taylor & Francis.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yani Huang, Richong Zhang, Zhijie Nie, Junfan Chen, and Xuefeng Zhang. 2025. A graph-based verification framework for fact-checking. *Preprint*, arXiv:2503.07282.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460, Online. Association for Computational Linguistics.

Yuqing Lan, Zhenghao Liu, Yu Gu, Xiaoyuan Yi, Xiaohua Li, Liner Yang, and Ge Yu. 2025. Multi-Evidence Based Fact Verification via A Confidential Graph Neural Network . *IEEE Transactions on Big Data*, 11(02):426–437.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. 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)*.

Vítor Lourenço, Aline Paes, and Tillman Weyde. 2025. Exploring content and social connections of fake news with explainable text and graph learning. In *Proceedings of the XXXV Brazilian Conference on Intelligent Systems (to appear)*. SBC.

Vítor Lourenço and Aline Paes. 2022. A modality-level explainable framework for misinformation checking in social networks. In *LatinX AI Workshop at 36th Conf. Neural Inf. Process. Syst., LXAI-NeurIPS*.

Mark Huasong Meng, Ruizhe Wang, Meng Xu, Chuan Yan, and Guangdong Bai. 2025. Detecting manipulated contents using knowledge-grounded inference.

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

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. 2024. Reuters institute digital news report 2024. Technical report, Reuters Institute for the Study of Journalism.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004. Association for Computational Linguistics.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Umair Qudus, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2025. Fact checking knowledge graphs – a survey. *ACM Comput. Surv.*

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 624–643. Association for Computational Linguistics.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop*.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving explainable fact-checking with claim-evidence correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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, Volume 1 (Long Papers)*.

Sebastián Valenzuela, Daniel Halpern, James E Katz, and Juan Pablo Miranda. 2019. The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, 7(6):802–823.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*.

Jinguang Wang, Shengsheng Qian, Jun Hu, Wenxiang Dong, Xudong Huang, and Richang Hong. 2025. End-to-end explainable fake news detection via evidence-claim variational causal inference. *ACM Trans. Inf. Syst.*, 43(4).

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, and Nikos Komninos. 2022. Evaluation of fake news detection with knowledge-enhanced language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1).

Zhouhui Wu, Nan Sun, Jiaojiao Jiang, and Shuiqiao Yang. 2025. Systematic approaches to fact verification: Evidence retrieval, veracity prediction, and beyond. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 271–289. Springer.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. Fire: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Cheng Xiong, Gengfeng Zheng, Xiao Ma, Chunlin Li, and Jiangfeng Zeng. 2025. Delphiagent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing & Management*, 62(6):104241.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836.

Xuan Zhang and Wei Gao. 2024. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

A Background

A.1 Key Concepts

Contrastive Explanations and Reasoning The principle of contrastive explanations is rooted on answering counterfactual why-question by comparing the actual outcome with hypothetical alternatives (Lipton, 1990; Guidotti, 2024; Verma et al., 2024). The contrastiveness presupposes that an explanation answers “Why did P happen?” in terms of “Why did P happen rather than Q?”, where P is an observed event and Q represents alternative hypotheses (Stepin et al., 2021). This approach ensures that explanations provide comprehensive information by distinguishing the chosen outcome from a set of contrastive hypothetical alternatives, establishing a minimum criterion where explanations must demonstrate why the observed event was more probable than its alternatives.

In short, contrastive explanations aim to answer why-questions by comparing an observed outcome P with counterfactual alternatives Q. Rather than just explaining why P occurred, they focus on why P happened instead of Q, highlighting why the chosen outcome is more plausible.

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

The use of LLMs to automate KG construction has been a theme of several works (Chen and Bertozzi, 2023; Zhang and Soh, 2024; Pan et al., 2024), specifically due to their performance on extracting entities and relations (Zhu et al., 2024).

A.2 Related Work

Contrastive Explanations and Reasoning Jacovi et al. (2021) applied contrastive explanations by projecting inputs into a latent space that captures only features distinguishing potential decisions, enabling models to better identify which aspects support or contradict specific predictions. Paranjape et al. (2021) extended this concept to commonsense reasoning tasks, showing that pre-trained language models can generate contrastive explanations that highlight key differences between alternatives, improving both performance and explanation faithfulness. These results show the effectiveness of a contrastive approach to reasoning and indicate its potential for claim verification.

Non-generative Automated Fact-checking AFC typically relies on text embeddings to represent claims and supporting documents. A notable example is dEFEND, which introduced a sentence-comment co-attention network to jointly analyse news content and user comments (Shu et al., 2019). Kotonya and Toni (2020) developed SBERT-FC, introducing the PUBHEALTH dataset to expand the thematic focus on public health and provide extended explainability analysis. GenFE and GenFE-MT (Atanasova et al., 2020) added joint optimisation of veracity prediction and explanation generation, showing improved classification and output quality.

CofCED (Yang et al., 2022) introduced a novel hierarchical architecture, combining an encoder with cascaded evidence selectors to process reports from multiple sources for claim verification and explanation generation. Whitehouse et al. (2022) integrated knowledge bases into pre-trained models and showed that using Wikidata improves accuracy, particularly for political claims where the knowledge base is timely and relevant. More recently, Qudus et al. (2025) surveyed the use of KGs for AFC.

Fact-checking Using LLMs The emergence of LLMs and their strong performance across domains has led to a new set of approaches that place them at the core of AFC. Whilst LLMs can give answers to factual questions, their reliability is limited by the extent of their training data and their tendency to hallucinate factual statements (Wang et al., 2024b; Augenstein et al., 2024). Several frameworks have been proposed to address the challenges of reliability and transparency in LLM-based fact-checking.

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

A popular approach to improving LLM fact-checking is Retrieval-Augmented Generation (RAG). Examples include the use of basic RAG-based architectures (Singal et al., 2024), RAG enhanced with fine-grained feedback mechanisms for optimising the retrieval task (Zhang and Gao, 2024), and more advanced RAG-based architectures focused on evidence retrieval and contrastive argument synthesis (Yue et al., 2024).

Other recent approaches focus on specific aspects of fact-checking. For instance, FIRE (Xie et al., 2025) proposes an iterative approach to fact-checking claims aiming to improve scalability and efficiency and Meng et al. (2025) address the challenge of zero-day manipulated content through real-time contextual information retrieval.

B Implementation Details

This section presents information about our experimental setup. We first detail the language models employed and their configuration settings, followed by the description and statistics of the datasets used in our evaluation (Section 4.1). Additionally, we present our modification to the *AlignScore* metric (Zha et al., 2023) (Section 4.3.1).

B.1 Models Settings

We evaluate our framework using a diverse set of language models, ranging from large-scale models available through Amazon Bedrock[3] to smaller, open-source alternatives from Hugging Face[4] (Hugging Face models were deployed using Amazon SageMaker AI[5]). Table 3 presents the complete list of models used in our experiments, along with their providers and custom hyperparameters. For all Amazon Bedrock models, we set the temperature to 0.0 to ensure deterministic outputs. For the Hugging Face models (SmolLM2 and Qwen3 families), we configure specific maximum length and token generation parameters to accommodate our fact-checking pipeline requirements.

B.2 Datasets

We evaluated the proposed approach on two fact-checking real-world datasets: LIAR-RAW and RAWFC (statistics shown in Table 4). LIAR-RAW (Yang et al., 2022) extends LIAR-PLUS (Alhindi et al., 2018) and contains 12,590 claims from Politifact[6] associated to six veracity classes. RAWFC (Yang et al., 2022) comprises of 2,012 claims collected from Snopes[7] with three veracity classes. Both datasets have their claims split into three sets: train, test, and validation, containing, respectively, 80%, 10%, and 10% of the claims. LIAR-RAW and RAWFC are used in Section 4.2 and Section 4.3.1. A modified version of LIAR-RAW, called *LAIR-RAW (2 labels)*, used in Sections 4.3.2 and 4.3.3, is also described in the table. This modification maps the original six labels into two classes: {`pants-fire`, `false`, `barely-true`} as `false`, and {`half-true`, `mostly-true`, `true`} as `true`, whilst preserving the original claims split.

[3] `https://aws.amazon.com/bedrock/`
[4] `https://huggingface.co/`
[5] `https://aws.amazon.com/sagemaker-ai/`
[6] `https://www.politifact.com/`
[7] `https://www.snopes.com/`

Table 4: Statistics of the LIAR-RAW and RAWFC datasets. $\mid \mathcal{C} \mid_{ALL}$ denotes the total number of claims, and $\mid \mathcal{R} \mid_{avg}$ represents the average number of reports per claim.

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

B.3 Weighted AlignScore and RQUGE

As described in Section 4.3.1, *AlignScore* is a metric based on a general function of information alignment to perform automatic factual consistency evaluation of text pairs (Zha et al., 2023). We extended AlignScore to weight its results by the distance of the predicted claim veracity class from the ground truth claim class.

For that, first, based on the semantic distance between classes, we progressively assign each class an integer numerical value ranging from 1 (y_{min}) to $\mid C \mid$ (y_{max}), where C is the set of possible labels. For instance, in the case of LIAR-RAW, we assigned each class an integer numerical value ranging from 1 (`pants-fire`) to 6 (`true`). After, we measure the mean squared error (MSE) between the predicted class. Finally, as depicted in Equation (2)), we weight the macro AlignScore for the pair answer summarisation text and original claim using the obtained MSE.

$$\text{AlignScore}_w = \left(1 - \frac{(\mathcal{V}_\mathcal{C} - y)^2}{(y_{\max} - y_{\min})^2}\right) \times \text{AlignScore}(A_\mathcal{C}, \mathcal{C}) \tag{2}$$

With this, we penalise the alignment scores of answer summary $A_\mathcal{C}$ and claim $\mathcal{C}$ where the fact-checking model prediction $\mathcal{V}_\mathcal{C}$ does not match the expected class y based on the answer summary,

Table 3: Overview of language models used in our experiments, including model providers and custom hyperparameter settings.

Model Name	Model Provider	Model Identifier	Custom Hyperparameters
Claude 3 Haiku	Amazon Bedrock	anthropic.claude-3-haiku-20240307-v1:0	`temperature = 0.0`
Claude 3.5 Haiku	Amazon Bedrock	us.anthropic.claude-3-5-haiku-20241022-v1:0	`temperature = 0.0`
Claude 3.5 Sonnet	Amazon Bedrock	us.anthropic.claude-3-5-sonnet-20241022-v2:0	`temperature = 0.0`
Claude 3.7 Sonnet	Amazon Bedrock	us.anthropic.claude-3-7-sonnet-20250219-v1:0	`temperature = 0.0`
Llama 3.3 70B	Amazon Bedrock	us.meta.llama3-3-70b-instruct-v1:0	`temperature = 0.0`
SmolLM2 135M	Hugging Face	HuggingFaceTB/SmolLM2-135M-Instruct	`max_length = 8192` `max_new_tokens = 128`
SmolLM2 1.7B	Hugging Face	HuggingFaceTB/SmolLM2-1.7B-Instruct	`max_length = 8192` `max_new_tokens = 128`
Qwen3 0.6B	Hugging Face	Qwen/Qwen3-0.6B	`max_length = 8192` `max_new_tokens = 32768`
Qwen3 1.7B	Hugging Face	Qwen/Qwen3-1.7B	`max_length = 8192` `max_new_tokens = 32768`

while considering that predicting closer classes is better than predicting distinct classes.

B.4 Weighted RQUGE

As described in Section 4.3.1, *RQUGE* scores the acceptability of a generated question given a context and an answer (Mohammadshahi et al., 2023). Let $q_i \in Q^{\mathrm{K}}_{ranked}$ be a (formulated contrastive) question for claim $\mathcal{C}$ (context) with corresponding answer $a_i \in \tilde{A}$; denote the macro score over a set of K questions as

$$\mathrm{RQUGE}_{\mathrm{macro}}(\mathcal{C}) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{RQUGE}(q_k, \mathcal{C}, a_k).$$

Analogously to Appendix B.3, we weight RQUGE by the distance between the model's predicted veracity class $\mathcal{V}_{\mathcal{C}}$ and the gold label y. Using the same class mapping $y_{\min} \to y_{\max}$ and mean-squared error term, the per-claim weight is

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

We define the weighted RQUGE for a single question and its macro form as:

$$\mathrm{RQUGE}_w(\mathcal{C}) = w_{\mathcal{C}} \times \mathrm{RQUGE}_{\mathrm{macro}}(\mathcal{C}). \quad (3)$$

This penalises question-quality scores when the verifier's prediction $\mathcal{V}_{\mathcal{C}}$ deviates from y, while granting higher credit to questions associated with *closer* (semantically adjacent) class predictions.

C Additional Experiments

C.1 Ablation Study

To evaluate the impact of KG-CRAFT's components and validate its effectiveness as a complete framework, we conducted a comprehensive ablation study on the LIAR-RAW and RAWFC datasets. Our analysis compares the full KG-CRAFT framework (proposed in Section 3 and results discussed in Section 4.2) against three key architectural variations: Naïve LLM, an LLM augmented with only the knowledge graph (KG), and an LLM that generates contrastive questions without using the KG structure (presented and discussed in Section 4.3.1). The results, presented in Table 5, provide a detailed breakdown of each component's contribution to the overall performance.

Naïve LLM The baselines Naïve LLM section (presented in Section 4.1 and results discussed in Section 4.2) shows the performance of aforementioned backbone LLMs when tasked with fact-checking using only the claim and its associated reports. This approach lacks any structured reasoning or pre-processing of the reports. The results for these models are notably lower than for KG-CRAFT, confirming the need for an enhanced reasoning mechanism.

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

LLM-based Contrastive Questions (No KG) This ablation directly addresses RQ2 (Section 4.3) by replacing the KG-based question formulation with questions generated purely by the LLM using a few-shot prompt (presented and discussed in Section 4.3.1). The results indicate that this approach is less effective than our KG-based method. The F1-scores are significantly lower, for instance, with Llama 3.3 70B achieving only a 29.68% F1-score on LIAR-RAW, compared to the 73.87% F1-score of the full KG-CRAFT framework. This suggests that LLMs, when prompted to generate their own contrastive questions, often fail to create questions that are both contextually relevant and aligned with the provided reports, resulting in overall lower information alignment (AlignScore) between the answer summaries and the original claims and overall lower question quality (RQUGE) between formulated questions, generated answers, and the original claims (Section 4.3.1). This finding reinforces the value of our knowledge graph approach for producing evidence-based questions.

KG-CRAFT (ours) The results of KG-CRAFT framework (proposed in Section 3 and results discussed in Section 4.2), which combines all proposed components, consistently outperform all other variations, achieving a new state-of-the-art on both datasets (Section 4.2). This confirms that the combination of knowledge graph extraction with the proposed contrastive reasoning significantly enhances LLMs' fact-checking abilities.

C.2 Evaluation with Other Datasets

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

C.2.1 Experimental Settings

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

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

Method	LIAR-RAW			RAWFC		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**
KG-CRAFT (ours)						
KG-CRAFT$_{C3.5}$	62.70	59.38	60.99	67.10	66.25	66.67
KG-CRAFT$_{C3.7}$	**73.92**	<u>70.86</u>	**72.36**	**69.37**	**68.59**	**68.98**
KG-CRAFT$_{L3.3}$	<u>77.38</u>	**70.67**	<u>73.87</u>	<u>81.63</u>	<u>81.53</u>	<u>81.58</u>
Naïve LLM						
Claude 3.5 Sonnet	29.25	27.83	28.52	54.85	55.04	54.94
Claude 3.7 Sonnet	28.26	26.63	27.42	55.90	57.12	56.50
Llama 3.3 70B	47.21	23.52	31.40	53.16	55.04	54.08
LLM with KG augmentation (no contrastive reasoning)						
Claude 3.5 Sonnet	39.96	37.52	38.70	56.73	56.51	56.62
Claude 3.7 Sonnet	60.70	56.42	58.48	68.27	66.09	67.16
Llama 3.3 70B	61.50	38.11	47.06	60.64	58.54	59.57
LLM-based contrastive questions (no KG)						
Claude 3.5 Sonnet	34.54	23.25	27.79	65.40	66.56	65.97
Claude 3.7 Sonnet	30.20	27.05	28.54	67.65	68.57	68.11
Llama 3.3 70B	42.18	22.90	29.68	64.98	64.99	64.98

Note: The <u>**best**</u> and **second best** results are highlighted across each dataset and metric. *KG-CRAFT* and *Naïve LLM* results are same in Table 1.

Comparisons The performance of KG-CRAFT is benchmarked against seven competing methods, which include encoder-based classifiers, graph models, program-style pipelines, and LLM-based approaches. MLA (RoBERTa) (Kruengkrai et al., 2021) proposes a sequence inference model which uses self-attention at both the token and sentence levels to capture information with a pre-LM encoder (RoBERTa). It also feeds static positional encodings into its multi-head attention mechanism. MULTIVERS (Wadden et al., 2022) leverages a multi-task learning approach aiming at multi-evidence scientific verification and rationale extraction from abstracts. ProgramFC (Pan et al., 2023) formulates fact verification as a programmatic pipeline with modular steps for evidence usage and decision making. PACAR (Zhao et al., 2024) is a prompting-based approach that aggregates evidence with consistency-oriented reasoning. GraphFC (Huang et al., 2025) encodes the report structure with an explicit graph and graph reasoning components. CO-GAT (ELECTRA) (Lan et al., 2025) applies graph attention over scientific evidence with a pre-LLM encoder (ELECTRA). GraphCheck (Chen et al., 2025) is a recent LLM-based verifier that incorporates lightweight graph signals and instruction-style prompting. We report KG-CRAFT with two backbones: KG-CRAFT$_{\mathrm{C3.5}}$ and KG-CRAFT$_{\mathrm{L3.3}}$, with the same settings as reported in our main experiments (Section 4.1).

Evaluation Metrics As in our main experiments (Section 4.1), we report precision (Pr), recall (Re), and F1-score (F1) results. In addition, to compare with GraphCheck, we report balanced accuracy (BAcc). For all metrics, higher values indicate better performance.

C.2.2 Results

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

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

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

Method	Pr	Re	F1
Previous Methods			
MLA (RoBERTa) (Kruengkrai et al., 2021)[a]	**80.62**	49.76	61.54
MULTIVERS (Wadden et al., 2022)	73.80	71.20	72.50
ProgramFC (Pan et al., 2023)[b]	-	-	71.82
PACAR (Zhao et al., 2024)	-	-	75.06
GraphFC (Huang et al., 2025)	-	-	**87.37**
CO-GAT (ELECTRA) (Lan et al., 2025)[d]	79.58	54.07	64.39
KG-CRAFT (ours)			
KG-CRAFT$_{C3.5}$	76.50	**75.18**	75.83
KG-CRAFT$_{L3.3}$	**84.58**	**81.53**	**83.03**

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

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

Method	PubHealth				SciFact			
	BAcc	Pr	Re	F1	BAcc	Pr	Re	F1
GraphCheck$_{L3.3}$ (Chen et al., 2025)	73.60	-	-	-	**89.40**	-	-	-
GraphCheck$_{Qwen\ 72B}$ (Chen et al., 2025)	71.70	-	-	-	86.40	-	-	-
KG-CRAFT (ours)								
KG-CRAFT$_{C3.5}$	61.02	76.50	75.18	75.83	75.17	76.50	75.18	75.83
KG-CRAFT$_{L3.3}$	**78.66**	72.82	73.89	73.35	81.52	84.58	81.53	83.03

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

strong overall results and cross-domain robustness evidenced by PubHealth.

D Prompt Engineering

This section presents the five prompts used in the scope of this work: Knowledge Graph Extraction (Section 3.1), Contrastive Question Answer Generation (Section 3.2), Answer Summarisation (Section 3.2), Claim Veracity Verification (Section 3.3), and Contrastive Question Formulation (Section 4.3.1). For each prompt, we provide the template.

D.1 Knowledge Graph Extraction Prompt

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

Knowledge Graph Extraction Prompt

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

D.2 Contrastive Question Answer Generation Prompt

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

Contrastive Question Answer Generation Prompt

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

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

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

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

## Output:
```

D.3 Answer Summarisation Prompt

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

Answer Summarisation Prompt

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

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

## Desired Outcome:
- A one paragraph summary of the
information contained in the question
and answer.
- Present the information neutrally,
without judging or labeling the
claim.
- Ensure that the summary is clear
and accurately based on the provided
context.
- Write in continuous prose (no lists,
bullet points, or meta-commentary).
```

```
- Do not add any utterances (for
example "Here are" statements) to the
final answer.
- Limit your answer to 200 words at
most.

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

## Output:
```

D.4 Claim Veracity Verification Prompt

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

Claim Veracity Verification Prompt

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

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

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

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

## Output:
```

D.5 Contrastive Question Formulation Prompt

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

Contrastive Question Formulation Prompt

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

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

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

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

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

## Input:
* Claim: {claim}
* Context: {reports}

## Output:
```