

Hallucination to Truth: A Review of Fact-Checking and Factuality Evaluation in Large Language Models

Subhey Sadi Rahman¹, Md. Adnanul Islam^{1,†}, Md. Mahbub Alam^{1,†},
 Musarrat Zeba^{1,†}, Md. Abdur Rahman¹, Sadia Sultana Chowdhury²,
 Mohaimenul Azam Khan Raiaan^{1,3}, Sami Azam^{3,*}

¹Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh

² Department of Computer Science and Engineering, Daffodil International University, Dhaka-1341, Bangladesh

³Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0909, Australia

[†] Equal Contributions.

* Corresponding Author: sami.azam@cdtu.edu.au

Abstract

Large Language Models (LLMs) are trained on vast and diverse internet corpora that often include inaccurate or misleading content. Consequently, LLMs can generate misinformation, making robust fact-checking essential. This review systematically analyzes how LLM-generated content is evaluated for factual accuracy by exploring key challenges such as hallucinations, dataset limitations, and the reliability of evaluation metrics. The review emphasizes the need for strong fact-checking frameworks that integrate advanced prompting strategies, domain-specific fine-tuning, and retrieval-augmented generation (RAG) methods. It proposes five research questions that guide the analysis of the recent literature from 2020 to 2025, focusing on evaluation methods and mitigation techniques. Instruction tuning, multi-agent reasoning, and RAG frameworks for external knowledge access are also reviewed. The key findings demonstrate the limitations of current metrics, the importance of validated external evidence, and the improvement of factual consistency through domain-specific customization. The review underscores the importance of building more accurate, understandable, and context-aware fact-checking. These insights contribute to the advancement of research toward more trustworthy models.

Keywords: Fact-checking, large language model, hallucination, LLM, retrieval augmented generation

1 Introduction

The growing use of Large Language Models (LLMs) in news, healthcare, education, and law means that their accuracy directly affects real-world decisions [1, 2]. These models often generate reliable information but may be false, which poses a risk of misinformation [3]. As newer fact-checking techniques, datasets, and benchmarks are published rapidly, it is challenging for researchers and practitioners to keep track of effective solutions [4]. This paper presents a systematic review of fact-checking related to LLMs, organizing recent advancements in this field, identifying and categorizing ongoing challenges, and highlighting potential avenues for future research. This is crucial as LLMs increasingly shape the way information is trusted and used in society.

1.1 Challenges

Fact-checking the output of LLM systems faces several intimidating challenges. The lack of standardized evaluation

metrics is one of the most notable challenges. Currently used ones quantify surface-level similarity rather than factual consistency [5, 6] and are therefore less effective at detecting nuanced errors.

Another key limitation of LLMs is hallucination. They tend to produce linguistically consistent but factually inaccurate or entirely fictional text [3]. This occurs due to language modeling and training on potentially stale or biased data using the probabilistic method [7, 8]. Dataset quality is also critical for fact-checking system performance. The majority of benchmarks either lack the realistic complexity of real-world claims and are domain-independent or are too narrow to be generalized. Furthermore, datasets with imbalanced classes can influence the model response and make the system less robust on a wide spectrum of topics [9].

1.2 Emerging Innovations

To address these issues, various innovations have been suggested, including Retrieval-Augmented Generation (RAG),

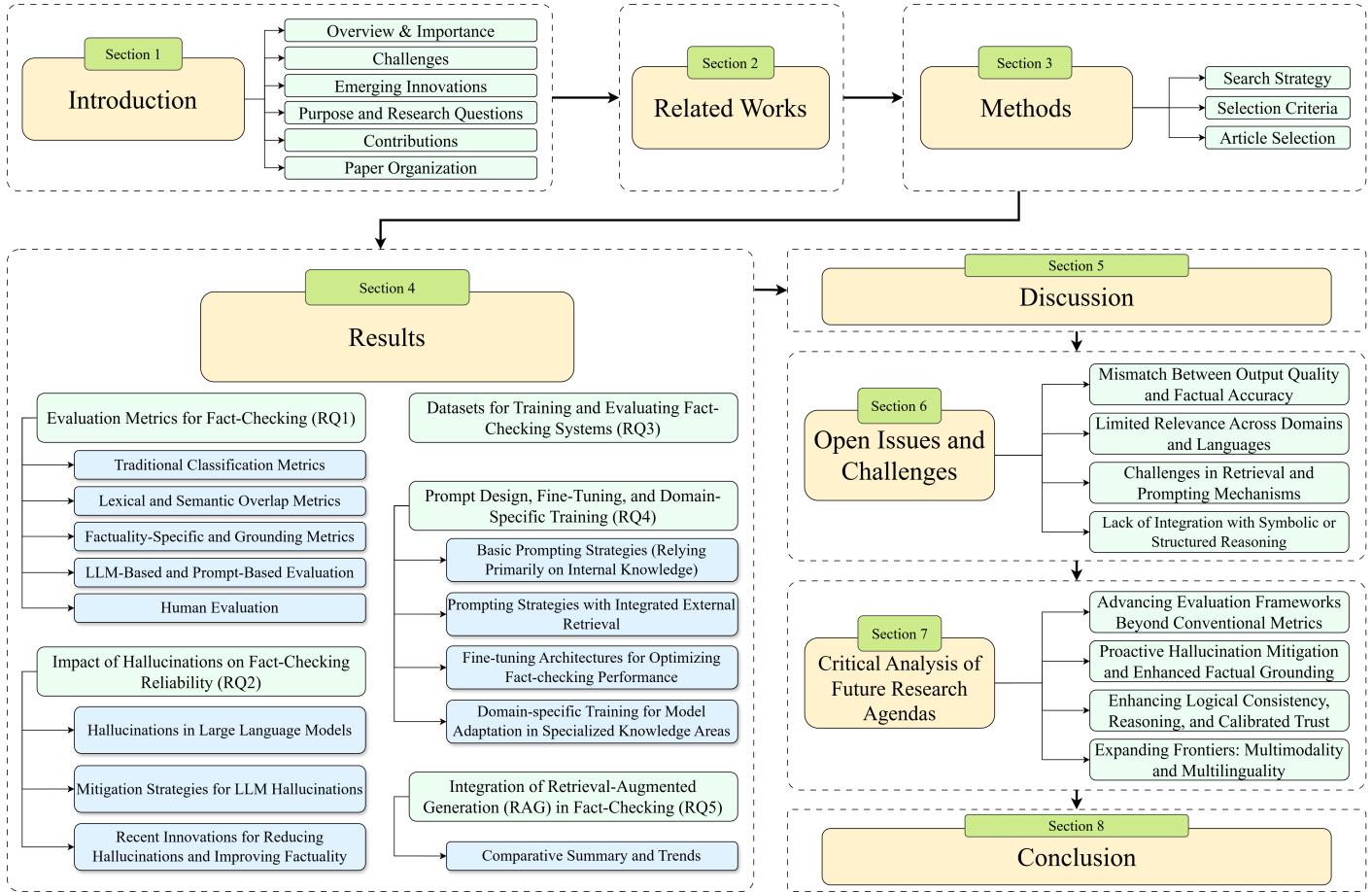


Figure 1: The fundamental content structure and categorization of this survey. The framework outlines eight sections, beginning with introduction, related works, and methods, followed by results addressing five research questions on evaluation metrics, hallucinations, datasets, prompting strategies, and retrieval-augmented generation. Subsequent sections cover discussion, open challenges, future research agendas, and conclusion, providing a coherent roadmap of the survey.

instruction tuning [10, 11], domain-specific fine-tuning [12], multi-agent systems [13, 14], automated self-correction and feedback mechanisms [15, 16], and integration with knowledge graphs [17, 18]. RAG stands out as a key technique that combines LLMs and external retrieval systems, aligning generated outputs with verifiable sources. RAG architectures have shown notable results in factuality and explainability [19] by allowing LLMs to access and cite external knowledge in real time. These methods are often augmented with advanced prompting techniques, such as hierarchical step-by-step reasoning and multi-agent collaboration [20, 21, 22].

1.3 Purpose and Research Questions

The objective of the review is to critically evaluate the current prospects of LLM-based fact-checking systems, identify key issues, and investigate the performance of existing solutions. Observing recent growing trends, this paper aims to

build more accurate, transparent, and scalable fact-checking systems in LLMs. The review is inspired by five fundamental research questions (RQ).

1. **RQ1:** What evaluation metrics are used to assess LLM-based fact-checking systems?
Rationale: To understand how system performance is measured and identify potential limitations or inconsistencies in current evaluation methods.
2. **RQ2:** How do hallucinations affect the reliability of LLM fact-checking?
Rationale: Hallucinations are caused by LLM due to the vast amount of training data containing refined and unverified information. Their output often contains hallucinated answers, which directly affect the trustworthiness and accuracy of the LLM fact-checking.
3. **RQ3:** What datasets are commonly used for training and evaluating fact-checking models?

Rationale: To assess the quality, coverage, and impact of the dataset on generalizability.

4. **RQ4:** How do prompting strategies and fine-tuning influence fact-checking performance?

Rationale: To analyze optimization techniques for LLMs in fact-checking contexts.

5. **RQ5:** How is RAG integrated into fact-checking?

Rationale: To evaluate the benefits and challenges of combining retrieval mechanisms with generative models.

1.4 Contributions

This paper presents three contributions to research on fact-checking using LLMs:

First. It offers a comprehensive taxonomy of evaluation metrics that categorizes widely used techniques by their methodological focus.

Second. The review combines a wide range of approaches to mitigate hallucinations in LLM output. These range from fine-tuning through domain-specific data to instruction tuning, adversarial training, and self-supervised feedback methods like Self-Checker [20]. In addition, multi-agent architecture and multi-step reasoning techniques are explored to enhance factuality and explainability.

Third. The paper offers novel insight into how dataset characteristics such as domain specificity, annotation quality, and multilingual coverage affect the performance of fact-checking systems [9].

1.5 Paper Organization

This paper is organized into eight sections: Section 2 reviews existing research on LLM-based fact-checking and highlights key gaps. Section 3 explains the methodology, including how the studies were selected and analyzed. Section 4 presents the findings based on our five core research questions. Section 5 discusses the implications, challenges, and limitations. Section 6 draws attention to open issues and challenges. Section 7 highlights the analysis of future research agendas, and Section 8 concludes the paper with key insights for building more accurate and reliable LLM-based fact-checking systems. Figure 1 illustrates the overall structure of the paper.

2 Related Works

LLM-generated texts are now widely used in various important sectors. Therefore, it is important to ensure their factual accuracy and reliability to maintain trust in these applications.

Several researchers have explored fact-checking methods in the context of LLMs. For example, Vykopal et al. [23] conduct a survey of approaches and techniques used in automated fact-checking using generative LLMs, such as claim detection, evidence retrieval, and fact verification. They introduce the concept of RAG, which can be used to mitigate challenges such as hallucinations and the use of out-of-date model knowledge, using external evidence. However, it does not address the effects of domain-specific training on LLM-based fact-checking, the challenges of RAG implementation, or how the quality of the dataset, the specificity of the domain, and the evaluation metrics influence the effectiveness of LLMs. Dmonte et al. [24] also explore LLM-based claim verification by analyzing full-system pipelines that include key stages such as evidence retrieval, prompt construction, and explanation generation. They review RAG techniques, including iterative retrieval and claim decomposition, which allow them to address issues such as hallucinations and the challenges of verifying complex or long claims. Evaluation metrics like FactScore and FEVER Score are highlighted to assess and improve factual accuracy. Similar to [23], this paper overlooks dataset quality, domain-specific challenges, and RAG implementation issues, which are significant in evaluating the reliability of LLM-based fact-checking.

In another work, Augenstein et al. [3] study the challenges of factual correctness in LLM. They focus on hallucinations, knowledge editing to reduce hallucinations, and the impact of misinformation that AI can spread, which also includes concerns about trust and misuse. They propose some mitigation strategies, such as RAG, although the discussion lacks depth on domain-specific RAG implementations and the associated challenges. Key evaluation metrics for LLM-based fact-checking systems are also discussed to assess factuality, consistency, and text quality, including TruthfulQA, FactScore, GPTScore, G-Eval, SelfCheckGPT, BERTScore, and MoverScore. However, the paper lacks a discussion of several critical technical aspects and challenges, including model interpretability, explainability, and practical implementation and integration of the proposed mitigation techniques in real-world settings. Wang et al. [25] offer a detailed survey of the factuality of LLM, providing a taxonomy of hallucination types and errors in both unimodal and multimodal tasks. A key contribution is mapping factuality challenges to algorithmic solutions and proposing improvements to factuality-aware model calibration. However, the paper does not discuss domain-specific challenges of RAG, prompt design, fact-checking component integration, or system-level architectures and deployment strategies for LLM verification environments.

Although most previous work has briefly discussed evaluation metrics, hallucination effects, and issues related to prompt- or fine-tuning methods, it has not examined the impact of datasets, particularly domain-specific ones. In addition, there is a wide gap between practical challenges

Table 1: Comparison of different papers concerning the key RQs. The following RQs guide our review: RQ1: Evaluation Metrics and Gaps; RQ2: Hallucination Effects and Mitigation; RQ3: Datasets and Impact; RQ4: Prompt and Fine-tuning; Domain-specific Training Effects; and RQ5: RAG and Domain-Specific Implementation Challenges.

| Paper | Metrics & Gaps | Hallucination & Mitigation | Datasets & Impact | Prompt, Fine-tuning & Domain Training | RAG & Domain | Key Notes |
|-----------------------|----------------|----------------------------|-------------------|---------------------------------------|--------------|--|
| Vykopal et al. [23] | ✗ | ✓ | ✗ | ✓ | ✗ | 1. Skips hallucination effects (RQ2) 2. Only mentions domain-specific fine-tuning (RQ4) 3. Only mentions RAG's impact (RQ5) |
| Dmonte et al. [24] | ✓ | ✓ | ✓ | ✓ | ✗ | 1. Skips domain-specific datasets (RQ3) 2. Only mentions domain-specific fine-tuning (RQ4) 3. Only mentions RAG's impact (RQ5) |
| Augenstein et al. [3] | ✓ | ✓ | ✓ | ✓ | ✗ | 1. Skips domain-specific datasets (RQ3) 2. Mentions prompt design in (RQ4) 3. Only mentions RAG's impact (RQ5) |
| Wang et al. [25] | ✓ | ✓ | ✓ | ✓ | ✗ | Only mentions RAG's impact (RQ5) |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | Fully addresses all RQs |

and considerations of RAG and domain-specific implementation. Most of the papers give RAG only a passing reference without discussing larger domain-specific problems. However, our work thoroughly addresses the primary research questions, especially the under-discussed areas, and thus provides a more detailed overview of the field. Table 1 presents a summary of the existing survey papers and shows how they relate to the key research questions of this study.

3 Methods

To explore how LLMs can be applied to fact-checking, we adopted a structured and practical approach inspired by well-established research methods [26]. The review process consisted of three key phases: (i) planning, (ii) data collection and analysis, and (iii) synthesis and reporting. **Firstly, we defined the overall scope of the review and ensured alignment with our RQs (i.e., RQ1-RQ5).** This included developing a detailed review protocol that specified the objectives, inclusion and exclusion criteria, and the databases to be searched. We also identified the dimensions most relevant to LLM fact-checking, including evaluation metrics, hallucinations, datasets, prompting and fine-tuning methods, and RAG, and set these as the guiding categories for further analysis.

We then performed a broad search across leading academic databases, using a combination of manual screening and automated tools to identify the most relevant and up-to-date studies. **Following this, we have analyzed the studies according to the methodological approach employed (i.e., benchmarking, prompting strategy, dataset evaluation), which RQ it addressed, and the contribution type (e.g., framework, metric, dataset,**

or application). At the end of our review, we brought together **these** insights from the selected studies to highlight what is already known, where the gaps are, and what future research should aim to address. By following this clear and step-by-step methodology, we have ensured that our findings are informative and reliable to advance the use of LLM in fact-checking tasks [27]. **The following sections detail the process in each aspect.**

3.1 Search Strategy

To ensure a comprehensive review of the relevant literature, we developed a focused search strategy employing well-defined keywords and Boolean operators. Our objective was to capture publications related to the application of LLMs in fact-checking and associated tasks. The search was conducted across several major academic databases, including Web of Science, arXiv, Scopus, and OpenReview, and covered a publication window from January 1, 2021, to September 15, 2025. For the Scopus and Web of Science databases, we utilized the following comprehensive search query:

```
("large language models" AND "fact-checking") OR
("LLM" AND "misinformation detection") OR ("automated fact verification" AND "LLMs") OR ("factuality evaluation" AND "natural language processing") OR ("hallucination" AND "LLMs") OR
("LLM hallucination") OR ("hallucination mitigation" AND "large language models") OR ("hallucination detection" AND "large language models") OR ("fact-checking datasets" OR "benchmark datasets for fact verification") OR ("retrieval-augmented generation" AND "fact-checking") OR
("RAG" AND "LLM") OR ("RAG" AND "fact-checking") OR ("fine-tuning" AND "fact verification models") OR ("prompt engineering" AND
```

("truthful generation" OR "fact-checking")) OR ("LLM-based fact verification" AND "NLP") OR ("misinformation detection" OR "fake news") AND "hallucination")

The specific search keys used for OpenReview and arXiv are detailed in our publicly available GitHub repository. Our selection process was designed to target significant research areas, including fact-checking methodologies, model assessment techniques, dataset analysis, and optimization procedures. The initial search produced the results summarized in Table 2.

Table 2: Number of Publications Retrieved from Each Database

| Database | Total Papers Retrieved |
|----------------|------------------------|
| Web of Science | 1,235 |
| arXiv | 9,180 |
| Scopus | 3,270 |
| OpenReview | ~1,000 |
| Total | 14,685 |

The high number of results from arXiv is primarily a consequence of its search limitations. The platform does not handle complex, multi-term queries well, which requires us to search for each keyword separately. This method inevitably led to the same articles being counted multiple times, explaining the inflated total.

3.2 Selection Criteria

Each article is carefully assessed using our evaluation criteria to determine whether it meets the inclusion or exclusion requirements. The key inclusion criteria (IC) and the exclusion criteria (EC) are shown in Figure 2.

3.3 Article Selection

We conducted a systematic review of articles from leading conferences and journals at the intersection of fact-checking, Natural Language Processing (NLP), and LLMs. The review focused on studies published between January 1, 2021, and September 15, 2025, that employed LLMs to verify external claims or factual content, excluding works that solely analyzed hallucinations or internal factual consistency.

From an initial pool of 14,685 records retrieved from various academic databases, we applied a multi-stage screening process, comprising duplicate removal, title and abstract screening, and full-text evaluation. The article selection workflow is illustrated in Figure 2, which outlines the stages of identification, screening, eligibility, and final inclusion. Finally, we selected 64 articles that meet our inclusion criteria and align with the objectives of this review, with publication dates ranging from September 2022 to August 2025.

An overview of the number of selected journals, conference proceedings, and preprints is shown in Figure 3. Figure 4 demonstrates the monthly publication frequency of the selected articles, with the majority published in 2024 and 2025. Furthermore, Figure 5 summarizes the geographical distribution of the publications, indicating that regions with well-established research ecosystems and institutional support contributed the largest share of works, which often correlates with higher representation in prestigious venues and greater citation visibility.

To ensure a comprehensive scope, we examined the distribution of publication venues. Due to the fact that various channels influence methodological priorities, venue analysis is crucial. RAG, prompting techniques, and assessment measures are commonly highlighted at prestigious NLP conferences, including ACL, EMNLP, and NeurIPS. Understanding publication frequency in conjunction with venue dispersion sheds light on the field's structural and temporal dynamics. Because it has a direct impact on methodological variety and research impact, venue concentration is important. Peer review procedures at high-impact conferences and publications are usually more stringent, guaranteeing methodological originality, more robust empirical validation, and wider recognition. Because of this, work that is published in these types of forums typically receives more attention and citations, which supports the prevailing paradigms in the area. This pattern also reflects how research environments with greater institutional support and global visibility tend to secure stronger representation in prestigious venues, which in turn shapes methodological diversity and amplifies citation impact. Domain-specific journals, on the other hand, provide important diversity by publishing contextually grounded research that may not meet the innovation-focused requirements of prestigious venues, despite the fact that they are frequently less referenced.

4 Findings from the Research Questions

In this section, we present a comprehensive key result finding focusing on evaluation metrics, the impact of hallucinations on LLMs, datasets, prompt design and fine-tuning, and integration of RAG.

4.1 Evaluation Metrics for Fact-Checking Systems (RQ1)

Evaluating LLMs for fact-checking and related areas such as grounded generation, summarization, and error detection is a crucial and evolving field. It addresses one of the most significant challenges in LLM deployment: their tendency to “hallucinate,” or generate text that sounds plausible but is factually incorrect [10, 18]. Evaluation in this context

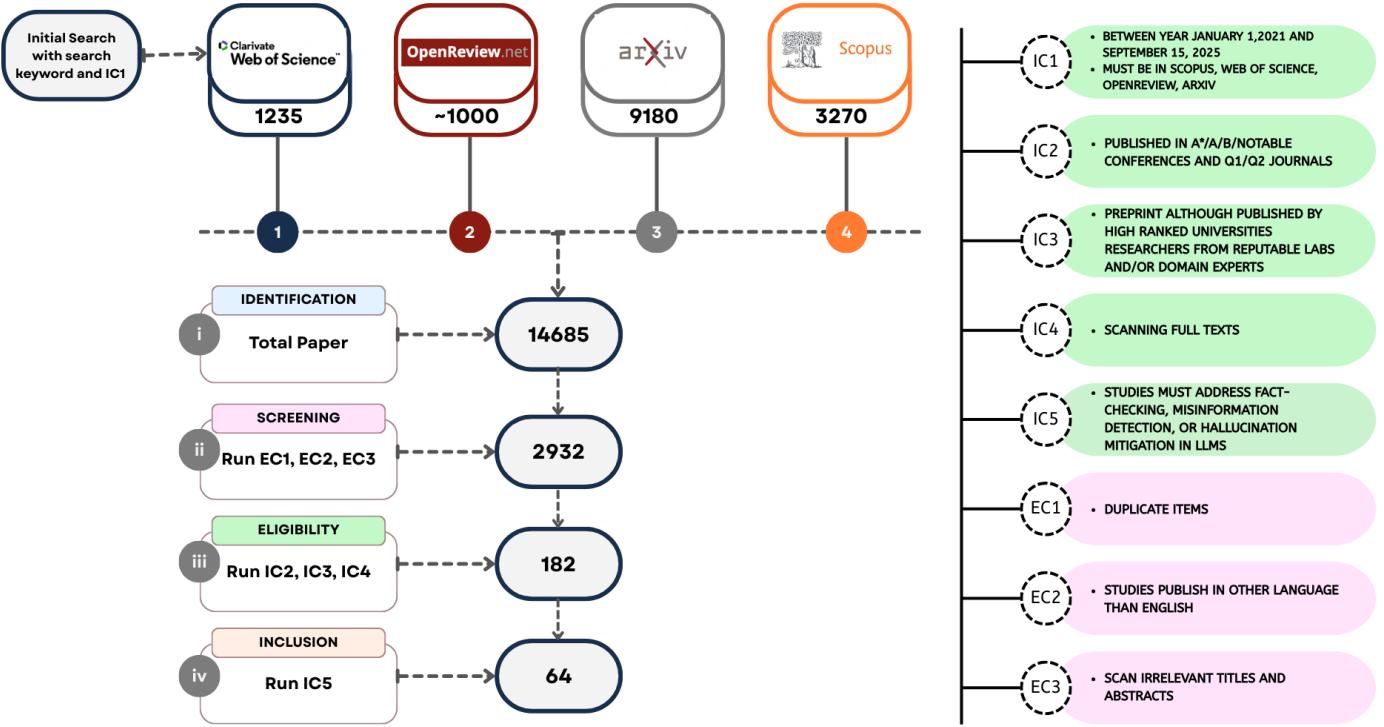


Figure 2: The article selection and screening process is on the left, and inclusion and exclusion criteria for article selection are on the right.

typically involves checking the model’s output against provided evidence or reliable external sources [10]. Previously, a wide range of methods have been developed, including using LLMs themselves as evaluators and building benchmarks that unify datasets and tasks [10]. Figure 6 illustrates a comprehensive summary of the complete set of metrics.

4.1.1 Traditional Classification Metrics

Most often, evaluation tasks are approached as classification problems, which determine whether a claim is true or identify errors in responses [28, 29]. Metrics like accuracy [28, 30, 31, 32, 33, 34], precision, recall, and F1-score [15, 21, 35, 36, 37, 38] are widely used for these tasks. In multiclass scenarios, such as classifying statements as supported, refuted, or inconclusive, macro-averaged versions of these metrics are employed [21, 28, 35]. These metrics also serve as standard measures in detection tasks [39]. For short-form responses, token-level precision with annotated answers is typical [15]. Token-level responses allow for partial evaluation, where an answer can be mostly correct, but as it is highly dependent on the specific tokenizer utilized, direct comparisons between different models are difficult.

These metrics offer quantitative performance indicators, making it easier to compare models or methods directly

[21, 28, 30, 32]. They often reduce complex output to a binary (i.e., correct or incorrect) judgment, overlooking reasoning quality or nuanced inaccuracies [18]. They may also be misleading in datasets with imbalanced labels [39].

4.1.2 Lexical and Semantic Overlap Metrics

When evaluating text generation tasks, such as summarization or dialogue, overlap-based metrics are commonly used. Lexical overlap metrics such as BLEU-4, METEOR, and chrF assess surface-level similarity [15]. ROUGE evaluates the extent to which summaries or explanations capture the core content [31, 40]. Semantic similarity metrics like BERTScore [3, 15, 34, 41], BLEURT [15], and cosine similarity measures [42] assess deeper semantics. For multimodal outputs, CLIP-Score compares image and caption embeddings to evaluate alignment [43].

These metrics are standard for assessing fluency and content similarity, and can capture meaning beyond exact word matches [15, 42, 43]. However, they do not measure the factual correctness. High overlap or semantic scores may still correspond to factually incorrect content. Older semantic metrics may not align well with the reasoning capabilities of modern LLMs [3, 40, 41].

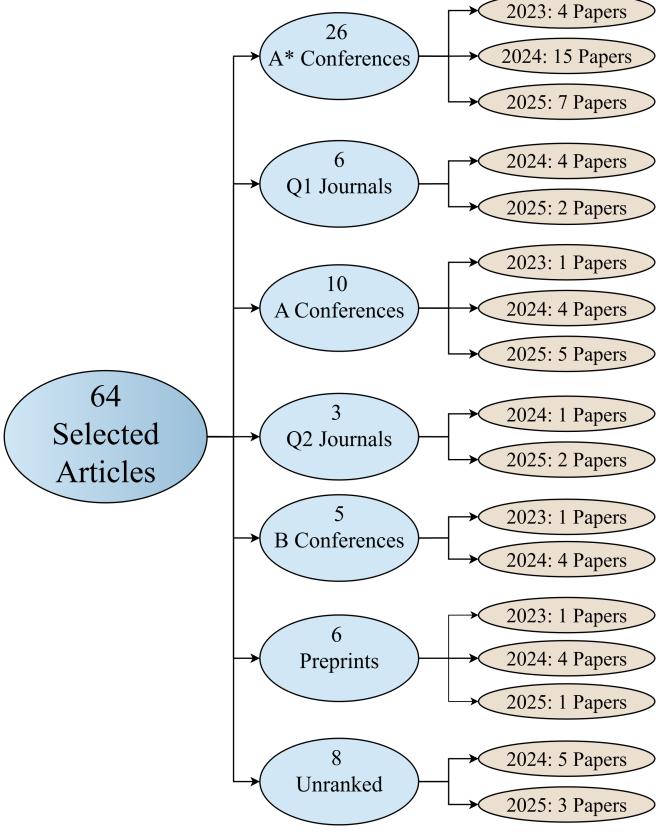


Figure 3: A visual summary of the articles selected from journals, conference proceedings, and preprints.

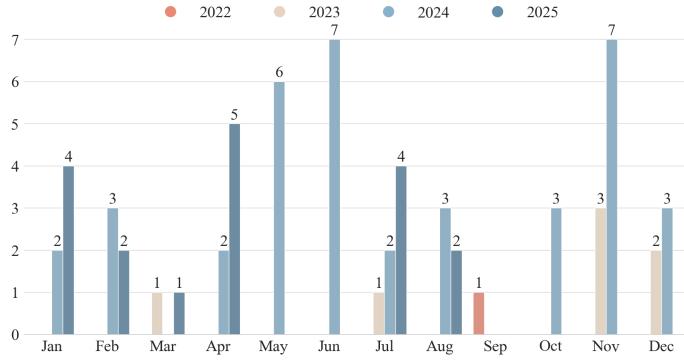


Figure 4: Grouped bar chart showing the monthly breakdown of publications from 2021 to 2025.

4.1.3 Factuality-Specific and Grounding Metrics

Specialized metrics have been developed to directly evaluate the factual consistency. These go beyond surface similarity and focus on whether the model’s claims align with the evidence. For example, benchmarks like LLM-AGGREFACT use detailed human annotations to assess support levels for claims [10]. It is a meta-benchmark for factuality that

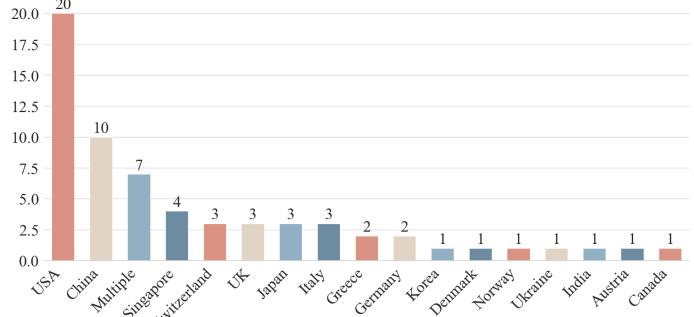


Figure 5: Bar chart illustrating the geographical breakdown of publications.

aggregates 11 different publicly available fact-checking and hallucination datasets. ReaLMistake focuses on binary error detection, especially in reasoning and context alignment [29]. Other tools like the LEAF fact-check Score compute the ratio of factually supported sentences to the total response [30] by decomposing each response into individual sentences, then each sentence is independently verified against retrieved external knowledge sources. Knowledge F1 (KF1), utilized by Peng et al. [15], measures the overlap between human-used and model-used knowledge. FactScore-Bio classifies responses based on retrieved evidence [44], and some methods aggregate multiple signals into a final factuality probability score [39]. Metrics such as Insight Mastery Rate (IMR) and Justification Flaw Rate (JFR) assess explanatory quality [45] where IMR represents the proportion of low-scoring fact-checking responses relative to the total number of questions, where a Grade of three or below (on a ten-point scale) indicates errors in the target LLM’s response. On the other hand, JFR denotes the percentage of cases where the target LLM conducted correct verdict prediction yet had poor justification, based on the conditions set by IMR. Natural Language Inference (NLI) techniques and textual entailment tasks also serve to classify claims as supported, refuted, or unverifiable [11, 33, 35, 46, 47].

Challenges include the complexity of strict entailment in language [48], potential metric bias [3], and reliance on high-quality annotated evidence. The Logical Consistency Matrix measures coherence under logical manipulations like negation or conjunction [17, 34]. Hit Rate (HR), used in evidence retrieval, tracks how often relevant documents are among the top results [49]. These methods are tailored for evaluating truthfulness and provide nuanced insights into factual grounding [10, 29, 30, 44].

4.1.4 LLM-Based and Prompt-Based Evaluation

A growing trend involves using LLMs themselves as evaluators. This includes having LLMs classify responses as correct or flawed, and rate the factual accuracy of claims when prompted [10, 29]. The LLM-as-a-judge paradigm

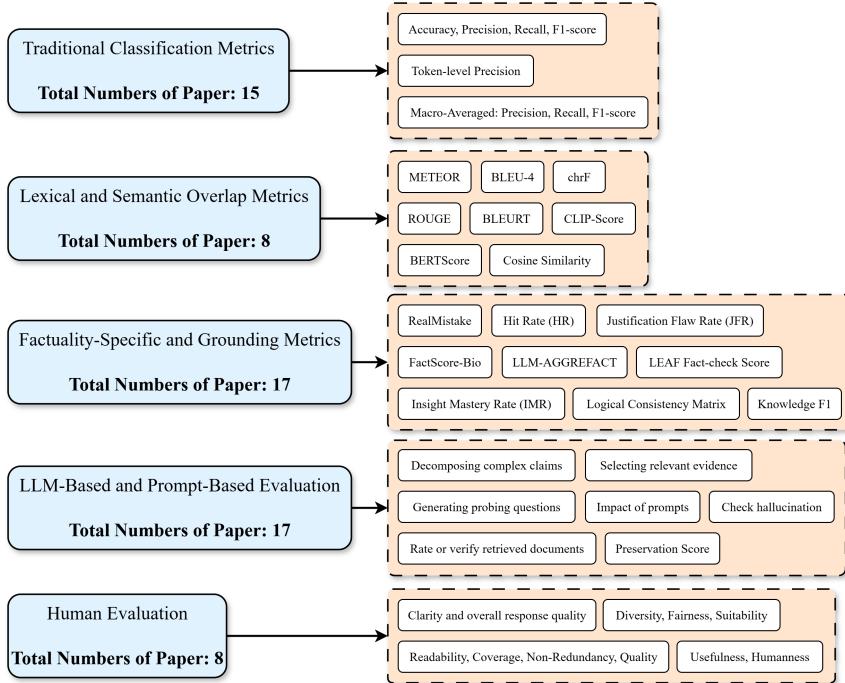


Figure 6: Taxonomy of evaluation metrics for fact-checking systems. The framework categorizes metrics into five groups. Each category is illustrated with representative measures and the number of papers adopting them, highlighting the breadth of methodologies employed in fact-checking research.

treats powerful language models as referees that compare and score the outputs of other models, which often produces results that closely align with human judgments [45, 50]. LLMs are also widely used for tasks such as decomposing complex claims [10, 42, 51], generating probing questions [30, 51, 52, 53], and selecting relevant evidence [20, 30, 42, 52]. While techniques like zero-shot, few-shot, Chain-of-Thought, ReAct, and HISS are not metrics themselves, their impact is assessed using factuality metrics [12, 21, 32, 54]. Some systems even use LLMs to rate and verify retrieved documents [51], or use them to check for hallucinations [15]. The Preservation Score evaluates how much original content remains intact after hallucination correction [49]. LLMs enable more nuanced and context-sensitive evaluations than traditional metrics [55]. They can reduce human effort in evaluation tasks as well [45]. Their performance can be inconsistent due to sensitivity to prompt phrasing, and they may introduce bias or misjudgments [29, 55].

4.1.5 Human Evaluation

Despite automation advances, human evaluation remains essential, especially for complex and subjective aspects like explanation clarity and overall response quality [56]. Evaluators often use Likert scales to rate Readability, Coverage, Non-Redundancy, and Quality [18, 21, 45]. In dialogue

tasks, several studies also assess aspects like Usefulness and Humanness [15]. Human-annotated data often forms the ground truth for many benchmarks [10, 29, 44], and evaluation criteria may include Redundancy, Diversity, Fairness, and Suitability [45]. Human judgments remain the gold standard, particularly for evaluating factual correctness and nuanced generation quality [18, 29], even though it is time-consuming, costly, and can introduce subjective variance depending on the evaluators and the criteria used [29, 45].

4.1.6 Comparative Summary and Trends

The landscape of LLM evaluation is becoming increasingly sophisticated. Traditional metrics like Accuracy and F1-score still serve as foundational tools for classification tasks [35, 21, 28]. However, more advanced evaluations focused on factuality and grounding are gaining prominence, especially in response to challenges such as hallucinations [10, 29, 30, 44].

Human-annotated benchmarks and specialized metrics help ensure robustness, while LLMs are now frequently integrated into the evaluation loop, whether to score, verify, or generate intermediate outputs [10, 29, 32]. Although promising, these LLM-based evaluations require careful validation against human judgments due to reliability concerns [29, 55]. Human evaluation continues to play a vital role, particularly in high-stakes and qualitative tasks. The

trend is toward hybrid frameworks that combine multiple evaluation strategies (e.g., automated metrics, LLM reasoning, and human oversight) to assess LLMs more holistically [15, 20, 51]. Thus, evaluating LLM fact-checking across diverse languages and modalities, including multi-modal or cross-lingual fact-checking, is an emerging frontier and demands the adaptation or creation of new evaluation techniques[22, 32, 48, 57].

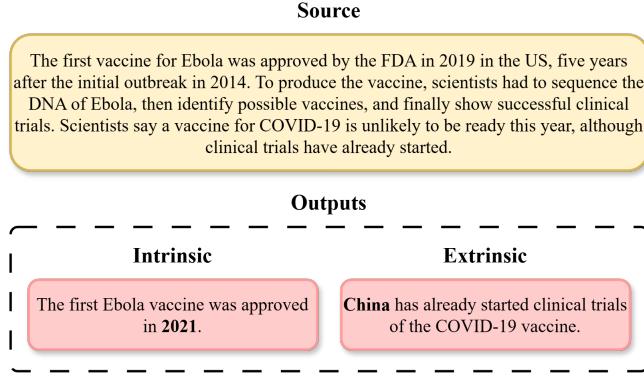


Figure 7: Intrinsic vs. extrinsic hallucinations in LLM outputs: The source text provides verifiable ground truth about Ebola and COVID-19 vaccines. The intrinsic hallucination example contradicts the fact explicitly stated in the source, whereas extrinsic hallucination introduces new information that is not supported by the source.

4.2 Impact of Hallucinations on Fact-Checking Reliability (RQ2)

Hallucinations in LLM refer to results that seem fluent, coherent, and linguistically correct but are factually inaccurate, nonsensical, unsupported, or completely fabricated [58, 59]. As these outputs are presented with the same level of confidence and linguistic fluency as factually accurate statements (see Figure 7), it is often difficult for users to detect them without external verification [58, 60]. These hallucinations can arise from contradictory, outdated, or misleading information in training corpora, as well as biases, lack of grounding, and even from prompts [3, 21, 61].

An overview of all the papers referenced in this section is presented in Table 3.

4.2.1 Hallucinations in LLMs

Nature and Types of Hallucinations. In the context of fact-checking, hallucinations manifest as intrinsic or extrinsic errors [62, 63, 64]. Intrinsic hallucinations occur when the model’s generated output contradicts the source content. In contrast, extrinsic hallucinations introduce information that cannot be verified by any provided evidence, resulting

in output unsupported by the source and often containing fabricated details [50, 60, 63].

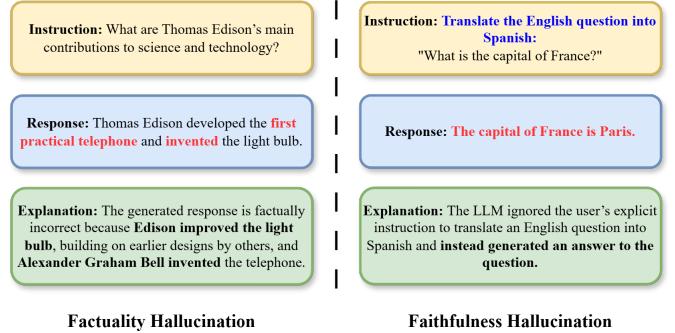


Figure 8: Two types of hallucination: Red-highlighted text shows hallucinated content, while blue-highlighted text reflects user instructions or context that conflict with the hallucination.

Summarization models may intrinsically hallucinate by stating a fact that contradicts the source article, or an open-ended question and answer (Q&A) model may extrinsically hallucinate entirely new (false) information that it presents as factual. The authors of [3, 49, 60] demonstrated that LLM hallucinations can generally be classified into two types (see example in Figure 8): (i) faithfulness, where the generated text does not accurately reflect the input context; and (ii) factuality, where the generated text is factually incorrect according to real-world knowledge.

Causes of Hallucination. Hallucinations arise from fundamental misalignment in the way LLMs are trained and used. The core training objective of most LLMs is to predict the next word in a sentence based on patterns learned from massive text data, not to guarantee truthfulness [34, 65, 66]. This means that models are optimized to produce text that is coherent and contextually appropriate rather than factually accurate [15, 17]. Thus, LLMs’ tendency to hallucinate can be traced to their optimization focus on linguistic fluency and coherence, rather than factual precision, especially when faced with queries outside of their training distribution or when internal knowledge conflicts arise [3]. During training, they absorb a large number of statements, including inaccuracies and fictional content, without an explicit mechanism to distinguish truth from falsehood [5, 6, 7, 8, 67]. As a result, an LLM may generate claims that sound confident and align with linguistic patterns in its memory, but are not grounded in facts. Wang et al. [44] demonstrate that a "knowledge error" can occur when the model produces hallucinated or inaccurate information due to a lack of relevant knowledge or internalizing false knowledge in the pretraining stage or the problematic alignment process.

Furthermore, since the LLM’s parametric knowledge

Table 3: An overview of studies on hallucinations covered in RQ2. The table organizes prior research by types and causes of hallucinations, their implications for reliability in fact-checking, and mitigation strategies. For each category, the publication years, representative authors, and study counts are reported.

| Topics Covered | Years | Authors | Count |
|---|------------------------------|---|-------|
| Nature and Types of Hallucinations | 2021, 2022, 2023, 2025 | Huang et al. [62], Ji et al. [63], Li et al. [64], Jing et al. [50], Huang et al. [60], Augenstein [3], Zhao et al. [49] | 7 |
| Causes of Hallucination | 2021, 2022, 2023, 2024, 2025 | Xie et al. [34], Zhou et al. [65], Wang et al. [66], Peng et al. [15], Ghosh et al. [17], Augenstein et al. [3], Lin et al. [5], Bender et al. [6], Paullada et al. [7], Ladhak et al. [8], Weidinger et al. [67], Wang et al. [44], Kasai et al. [19], Tran et al. [30], Cheung et al. [46], Li et al. [68], Onoe et al. [69], Tang et al. [10], Yao et al. [70] | 19 |
| Implications for Reliability in Fact-Checking | 2023, 2024, 2025 | Peng et al. [15], Si et al. [33], Li et al. [71], Zhao et al. [49], Xie et al. [34], Quelle et al. [48], Hu et al. [12], DeVerna et al. [72], Singhal et al. [53], Zhao et al. [14], Augenstein et al. [3], Wang et al. [44], Jing et al. [50] | 13 |
| Fine-tuning and Instruction Tuning | 2021, 2023, 2024, 2025 | Tang et al. [10], Setty et al. [11], Hu et al. [12], Zhang et al. [28], Zhao et al. [14], Cheung et al. [46], Tran et al. [30], Qi et al. [31], Luo et al. [73], Leite et al. [55], Jing et al. [50] | 11 |
| RAG | 2023, 2024, 2025 | Singhal et al. [53], Quelle et al. [48], Augenstein et al. [3], Si et al. [33], Peng et al. [15], Sankararaman et al. [39], Khalil et al. [22], Zhang et al. [51], Xie et al. [34], Tran et al. [30], Wei et al. [74], Zhao et al. [14], Qi et al. [31], Li et al. [20], Giarellis et al. [18], Ma et al. [16], Zhao et al. [49], Li et al. [71], Ghosh et al. [17], Tang et al. [10] | 20 |
| Adversarial Tuning | 2025 | Leippold et al. [75] | 1 |
| Automated Feedback Mechanisms and Self-Correction | 2023, 2024, 2025 | Peng et al. [15], Ma et al. [16], Xie et al. [34], Tran et al. [30], Fadeeva et al. [76], Ghosh et al. [17], Ge et al. [43], Zhao et al. [49] | 8 |
| Hybrid Approaches and Multi-Agent Systems | 2023, 2024, 2025 | Zhang et al. [21], Zhao et al. [14], Ma et al. [16], Kupershtain et al. [13], Li et al. [71], Giarellis et al. [18], Hu et al. [12], Ghosh et al. [17], Jing et al. [50] | 9 |
| Multimodal Fact-Checking | 2024 | Cao et al. [36], Ge et al. [43], Yao et al. [77], Papadopoulos et al. [78], Sharma et al. [79], Qi et al. [31], Wang et al. [80], Kakizaki et al. [81], Geng et al. [57], Khalil et al. [22] | 10 |
| Multilingual Fact-Checking | 2024 | Quelle et al. [48], Zhang et al. [82], Shafayat et al. [83], Siino et al. [84], Shcharbakova et al. [85], Jannah et al. [86] | 6 |
| Domain-Specific Fact-Checking | 2023, 2024, 2025 | Zhang et al. [28], Tran et al. [30], Vladika et al. [38], Zhao et al. [49], Xiong et al. [87], Jing et al. [50], Chatrath et al. [54], Khalil et al. [22], Choi et al. [47], Zhang et al. [21], Hu et al. [12], Leite et al. [55], Wang et al. [44], Qi et al. [31], Choi et al. [88], Pisarevskaya et al. [41], Liu et al. [37] | 17 |
| Enhancing Explainability and Trust | 2023, 2024, 2025 | Ding et al. [56], Sankararaman et al. [39], Quelle et al. [48], Zhao et al. [14], Qi et al. [31], Vladika et al. [38], Krishnamurthy et al. [42], Ghosh et al. [17], Leite et al. [55], Giarellis et al. [18] | 10 |
| Hierarchical Prompting and Multi-Step Reasoning | 2023, 2024 | Zhang et al. [21], Khalil et al. [22], Zhao et al. [14] | 3 |

base is fixed after training, it can become outdated or insufficient, leading to guesses on topics that it does not know [19, 30, 46, 68, 69]. When prompted “closed-book” (without access to external data), a model faced with an unknown fact will often fill the gap by generating a plausible answer, which is a major source of hallucination [10]. In other cases, even if grounding documents are provided, the model might improperly blend or misattribute information from these sources, causing intrinsic hallucinations [10]. Hallucinations may also be an inherent adversarial vulnerability of LLMs. Even nonsense or out-of-distribution prompts can trigger the model to produce a false but fluent response [70]. This suggests that, beyond knowledge gaps, the model’s sensitivity to input perturbations and over-reliance on spurious

correlations in training data can induce hallucinations.

Implications for Reliability in Fact-Checking. Hallucinations directly undermine the reliability of LLMs as fact-checkers across domains, and in some cases, they can be fatal when deployed for mission-critical tasks [15, 33, 49, 71]. Studies estimate that even advanced models like GPT-4 and LLaMA-2 produce inaccurate factual statements in approximately 5–10% of their responses to general knowledge queries [34]. When LLMs are used in fact-checking systems, hallucinations can lead to misclassification of claims, which can lead to an underestimation of the accuracy and trustworthiness of fact-checking procedures [48]. Moreover,

hallucinations risk amplifying misinformation when LLM-generated content aligns with existing false narratives, unknowingly facilitating their widespread acceptance in the eyes of the public [12, 53, 72].

The complexity of detecting hallucinated content, particularly when intertwined with accurate information, complicates the verification process and increases the mental workload for human fact checkers, potentially affecting decision-making, as evidenced by studies that indicate reduced discernment after LLM-assisted fact-checking [33, 72]. Additionally, automated fact-checking systems comprising LLMs for evidence retrieval, claim interpretation, or verdict generation are highly vulnerable, as hallucinations at any stage can contaminate the entire fact-checking pipeline, ultimately resulting in unreliable outcomes [3, 14, 44]. Recent work emphasizes the need to quantify these phenomena systematically to enhance model reliability, proposing methodologies specifically aimed at measuring and addressing hallucination severity within faithfulness evaluations [50]. Singhal et al. [53] state that, in terms of fact-checking, simply assigning a veracity label is inadequate. The prediction must be supported by evidence to ensure the system’s transparency and to bolster public trust.

4.2.2 Mitigation Strategies for LLM Hallucinations

Fine-tuning and Instruction Tuning. Fine-tuning pre-trained LLMs on datasets tailored to a specific domain or task, emphasizing factuality, substantially enhances their reliability [10]. Quantitative comparisons supporting this observation are summarized in Table 7. Table 4 illustrates different fact-checking frameworks and their demonstrated approach to mitigate hallucination. Instruction tuning, a specialized variant of fine-tuning, trains models to better adhere to explicit instructions aimed at factual and verifiable responses. Specifically, domain-specific adaptation involves fine-tuning on datasets rich in factual claims and evidence from specialized areas, such as news, medical, political, religious, or legal domains, thus helping models learn the intricacies of factual language and reasoning pertinent to these fields [11]. Tang et al. [10] showed that fine-tuning transformer-based models, RoBERTa, DeBERTa, and T5 models on synthetic and ANLI data boosts robustness and improves performance. Hu et al. [12] proposed that utilizing LLMs’ ability to provide rationales could enhance fine-tuned small language models (SLMs), thereby improving fake news detection performance. Zhang et al. [28] demonstrated that in clinical claim evaluation, domain-tuned discriminative models such as BioBERT with 80.2% accuracy outperformed both zero-shot and fine-tuned generative LLMs like Llama3-70B, even after tuning. Zhao et al. [14] fine-tuned Pretrained LMs utilizing various strategies such as BERT-FC with dual loss, LIST5 with list-wise reasoning, and T5

language model, RoBERTa using NLI, and MULTIVERS with multitask learning.

In terms of instruction-following techniques for factuality, several notable works, including Cheung et al. [46], explicitly enhanced instruction-following models by integrating external knowledge specifically for fact-checking. They illustrated that targeted fine-tuning can significantly improve factual accuracy. Setty et al. [11] demonstrated that smaller, finely-tuned models can sometimes surpass larger, more general models in accuracy for fact-checking tasks. Tran et al. [30] introduced two parallel self-training methods to update LLMs and boost factual reliability: Supervised Fine-Tuning (SFT) using verified responses, and Simple Preference Optimization (SimPO) using fact-based ranking. Qi et al. [31] proposed a two-stage instruction tuning method to adapt InstructBLIP for OOC misinformation detection, first aligning it to the news domain using NewsCLIPpings data [73] and then fine-tuning it on GPT-4-generated inconsistency explanations. Leite et al. [55] introduced a multi-stage weak supervision approach utilizing instruction-tuned LLMs prompted with 18 credibility signals to generate weak labels, which are then aggregated to predict content veracity, thereby minimizing hallucinations and enhancing transparency for human fact checkers. Jing et al. [50] fine-tuned two HHEM DeBERTa NLI models on synthetic data and found that the HHEM model with fine-tuning on synthetic data can outperform LLMs in domain-specific evaluation, given carefully crafted training data.

Retrieval-Augmented Generation. RAG has emerged as a prominent technique to ground LLM outputs in external, verifiable knowledge sources, with the aim of improving the factual context, reducing hallucination, and reliance on potentially flawed internal knowledge [3, 33, 48, 53, 89]. RAG architectures generally include a retriever module that gathers relevant information from external sources, such as web documents or databases, and a generator module (the LLM) that synthesizes this retrieved information and formulates answers or assessments [15, 39].

Several studies have demonstrated the effectiveness and variations of RAG. For instance, Singhal et al. [53] reported how integrating external knowledge and feedback loops can significantly improve factuality. Table 7 illustrates these performance gains across multiple benchmarks, including PolitiFact and PubMedQA. Peng et al. [15] introduced LLM-AUGMENTER that enhances LLM responses by retrieving external knowledge, linking raw evidence with related context, verifying outputs, and iteratively refining prompts using automated feedback until the response is free from hallucination and factually grounded. On the other hand, Quelle et al. [48] allowed LLMs to perform Google searches to gather evidence related to the claim. Sankararaman et al. [39] presented a method, Provenance, where a dual-stage cross-encoder framework, one stage filters and weights context items, and the other assesses the factuality

Table 4: An overview of notable fact-checking frameworks for LLMs and their approach to mitigate hallucination.

| Framework | Hallucination Types Addressed |
|----------------------------------|--|
| MiniCheck | Specifically targets factual errors where LLM-generated text is not grounded in or is inconsistent with provided source documents. It is designed to handle complex errors that require synthesizing information across multiple sentences and verifying multiple facts within a single claim. |
| CliniFact | Addresses factual errors in the highly specialized medical domain by verifying clinical research claims. It also targets errors in Logical Reasoning, as it evaluates logical statements derived from hypothesis testing in clinical trials. |
| VisualFactChecker | Explicitly designed to mitigate hallucinations in image captioning, such as generating descriptions of non-existent objects (content) or incorrect attributes (shape, color), by using visual tools (object detection, VQA) for verification. |
| FACT-GPT | Addresses the challenge of recurring misinformation by matching new social media posts to previously debunked claims. It tackles variations of false claims that entail or contradict a known falsehood. |
| Hierarchical Step-by-Step (HiSS) | This method uses a prompting technique to mitigate hallucination, where prompts are given to LLMs to do fine-grained checking of news claims. Here, prompts are given to the LLM to decompose the claim into subclaims and verify each subclaim step-by-step by raising and answering a series of questions. For each question, a prompt is again given to the LLM to assess if it is confident to answer it or not, and if not, the question is given to a web search engine. The search results are then inserted back into the ongoing prompt to continue the verification process. |
| LLM-KG Framework | Mitigates hallucinations by grounding LLM responses in a structured and verifiable Knowledge Graph. It helps prevent the generation of plausible but incorrect facts by providing explicit, verified triplets as context. |
| LLM-AUGMENTER | Mitigates hallucinations by augmenting a fixed LLM with external knowledge to generate grounded responses. It uses automated feedback to iteratively revise responses and improve their factuality score against the evidence. |
| OpenFactCheck | Evaluates and identifies various types of factual errors, including Knowledge Errors (hallucinated/inaccurate information), Over-commitment Errors (failing to recognize false premises), and Disability Errors (outdated information). |
| MEDICO | This is a framework specifically designed for hallucination detection and correction. It fuses evidence from multiple sources (web, knowledge bases, knowledge graphs) to detect factual errors, provide a rationale, and iteratively revise the hallucinated content. |
| Yours Truly | This framework is designed to fact-check social media claims by breaking down compound sentences into atomic claims and verifying each against a real-time database of fact-checked articles. It addresses the challenge of hallucinations in LLMs by verifying outputs against an external, curated knowledge source (FactStore). |
| Logical Consistency Framework | This framework addresses inconsistencies in LLM responses, which are a vulnerability related to hallucination. It assesses and improves the logical consistency of LLMs when performing fact-checking against Knowledge Graphs on complex queries involving logical operators, aiming to make the LLM less likely to hallucinate by ensuring its reasoning is logically sound. |

score, with the scores aggregated for threshold-independent evaluation. Khaliq et al. [22] applied multimodal reasoning within RAG, proposing Chain of RAG (CoRAG) and Tree of RAG (ToRAG), specifically for political contexts.

Reinforcement learning has also been used in the RAG process. For example, Zhang et al. [51] investigated reinforcement learning to optimize retrieval processes. [34] proposed an iterative retrieval and verification mechanism to refine accuracy further. Singhal et al. [53] introduced a RAG-based fact-checking system where the core pipeline retrieves top-3 documents with FAISS, extracts evidence, and then uses the evidence in classification, combining the In-Context Learning (ICL) capabilities of multiple LLMs, achieving a 22% gain on the Averitec dataset. Tran et al. [30] introduced Fact-Check-Then-RAG, which improves the factual accuracy of LLM by evaluating each fact with SAFE

[74], retrieving relevant data using ColBERT from MedRAG for failed checks, and regenerating responses via a RAG-enhanced prompt. Zhao et al. [14] introduced a multi-agent and retrieval-augmented fact-checking system that dynamically selects reasoning tools and external evidence to handle diverse multi-hop verification tasks.

To improve out-of-context (OOC) detection, Qi et al. [31] integrate Google’s Entity Detection API for visual grounding and perform external verification through LLM to verify news captions against evidence retrieved from reverse image searches. Li et al. [20] introduced SELF-CHECKER, which verifies input by extracting simple claims, generating search queries on various external knowledge sources, e.g., Bing Search API, retrieving evidence from knowledge sources, e.g., Wikipedia, Reddit messages, and predicting each claim’s veracity based on its selected

evidence sentences. Giarelis et al. [18] proposed a unified LLM-KG framework for fact-checking, which retrieves relevant facts from Knowledge Graphs (KGs) and injects them into the LLM prompt for response generation. Ma et al. [16] introduced a Logical and Causal fact-checking method (LoCal), a LLM-driven multi-agent framework that breaks down complex claims, resolves them through specialized reasoning, and validates consistency using logical and counterfactual evaluators in an iterative process. Zhao et al. [49] introduced MEDICO, where the system retrieves evidence from various sources, search engines, knowledge bases (KB), KGs, and user files, then re-ranks and fuses it by concatenation or Llama3-8B-based summarization. Li et al. [71] introduced FactAgent that uses LLM's internal knowledge (i.e., Phrase, Language, Commonsense, Standing tools) and another that integrates external knowledge tools (i.e., URL and Search tools - SerpAPI). Ghosh et al. [17] introduced LLMQuery that evaluates LLMs' logical consistency in fact-checking by retrieving subgraphs from KGs using BFS or ANN-based vector embedding methods, ensuring concise and relevant context is fed to the model. Finally, Tang et al. [10] introduce an efficient method specifically tailored to evaluate claims generated by LLM against grounding documents, which forms a core component of a comprehensive evaluation of RAG.

Adversarial Tuning. Adversarial training presents LLMs with specifically designed examples intended to uncover hallucinations or factual errors, thus training the models to recognize and handle these challenging inputs accurately and improve their robustness against generating misinformation. The study by Leippold et al. [75] introduced CLIMINATOR, an acronym for CLImate Mediator for INformed Analysis and Transparent Objective Reasoning, and an AI-based tool utilizing a Mediator and adversarial Advocate framework to automate climate claim verification by simulating structured debates, including climate denial perspectives, iteratively reconciling diverse viewpoints to converge towards scientific consensus consistently and thus improving accuracy and reliability.

Automated Feedback Mechanisms and Self-Correction. Incorporating automated feedback loops and enabling LLMs to self-critique and correct their outputs are emerging as powerful strategies, exemplified by several key approaches. Peng et al. [15] proposed an automated feedback mechanism, LLM-AUGMENTER, which substantially reduces ChatGPT's hallucinations without sacrificing the fluency and informativeness of its responses by iteratively revising LLM prompts to improve model responses using feedback generated by utility functions, e.g., the factuality score of a LLM-generated response. Ma et al. [16] utilized two evaluating agents that iteratively reject or accept solutions and trigger new decomposition or reasoning rounds until consistency is

achieved. Additionally, iterative refinement processes, such as those demonstrated by Xie et al. [34], continuously check and improve model outputs through multiple verification cycles; Tran et al. [30] introduced LEAF, a self-training loop that utilizes fact-check scores as automated feedback.

Furthermore, the uncertainty quantification approach proposes using token-level uncertainty measures to detect potential hallucinations and trigger additional verification steps, thereby enhancing overall reliability and accuracy [76]. Ge et al. [43] also showed in their framework that LLMs utilize object-detector and VQA results to mitigate hallucinated contents automatically and fact-check proposed captions. In the work of Zhao et al. [49], it iteratively corrects only hallucinated parts in generated content using Chain-of-Thought (CoT) prompting, while enforcing minimal edits via Levenshtein-based preservation scoring.

Hybrid Approaches and Multi-Agent Systems. Combining multiple strategies or employing multi-agent architectures, wherein different LLM agents handle specialized sub-tasks within the fact-checking process, has emerged as a growing trend, exemplified by hierarchical prompting and planning methods such as the work by Zhang et al. [21] systematically guided models through complex claim verification, and Zhao et al. [14] utilized LLMs for structured planning and reasoning tasks.

Additionally, multi-agent systems have gained attention, as seen in studies like LoCal [16], employing multiple specialized LLM agents (decomposer + reasoner + two evaluators) to address logical and causal dimensions of fact-checking. Further explored by Kupershtain et al. [13] and Li et al. [71], both of which investigate the capabilities of agent-based frameworks specifically for fake news detection. Li et al. [71] also introduced FactAgent that behaves in an agentic manner, emulating human expert behavior utilizing several tools (Phrase, Language, Commonsense, Standing, URL, and Search tools) around a single LLM.

Moreover, the integration of structured knowledge sources is further explored, advocating the combination of LLMs with knowledge graphs to comprise structured factual information [18]. Additionally, Hu et al. [12] designed a novel approach, ARG and its distilled version ARG-D, that complements small and large LMMs by selectively acquiring insights from LLM-generated rationales for SLMs. This combination of LLM+SLM has shown superiority over existing SLM/LLM-only methods. By introducing a hybrid KG retrieval + LLM generation approach and supervised fine-tuning, Ghosh et al. [17] improved the logical consistency of LLMs on the complex fact-checking task. Similarly, Jing et al. [50] combined rubric-prompted LLM judges and an NLI cross-encoder (HHEM) in the same framework.

4.2.3 Recent Innovations for Reducing Hallucinations and Improving Factuality

Beyond the above core strategies, recent studies have further tried to address hallucination issues and ensure LLMs remain faithful to facts. These methods range from innovative prompting techniques and multi-step reasoning procedures to incorporating multiple modalities, as well as building self-checking mechanisms and uncertainty estimates into LLM responses. We highlight several promising directions below, along with their advances in architecture, evaluation, and practical deployment.

Multimodal Fact-Checking. Misinformation is increasingly multimodal, combining text, images, and videos, thus addressing hallucinations and ensuring factuality in LLMs processing such data has become crucial. Efforts to integrate visual and textual evidence include the study by Cao et al. [36], which employed graph attention networks to consolidate multimodal knowledge for verifying claims, and Ge et al. [43], focusing specifically on accurate captioning of images utilizing visual fact-checking through object detection and VQA models. Then, to address multimodal fact-checking with explanation generation, Yao et al. [77] presented Mocheg, a large-scale benchmark dataset, and an initial demonstration of existing methods' performance, which generated unsatisfactory results. Subsequently, Papadopoulos et al. [78] proposed the RED-DOT model, integrating a module called "relevant evidence detection" (RED) that can evaluate the relevancy of each piece of evidence and achieves up to 33.7% accuracy gain on the VERITE benchmark. Furthermore, Sharma et al. [79] evaluated the visual grounding capabilities inherent in language models, while [31] introduced a multimodal LLM-SNIFFER which analyzes both the consistency of the image-text content and the claim-evidence relevance using InstructBLIP and GPT-4V. Fake news detection is an emerging and vital domain addressed by Wang et al. [80], where they proposed a hybrid model called FND-LLM that combines SLMs and LLMs. This results in significant accuracy improvements of 0.7%, 6.8%, and 1.3% on the Weibo [90], Gossipcop [12], and Politifact datasets. Recently, Kakizaki et al. [81] introduced a system called Multimodal Automated Fact-checking via Textualization (MAFT), that textualizes content such as images, video, and audio for LLM analysis, which generates comprehensive and interpretable reports that explain the verification steps to combat misinformation. Practical applications are explored in [57], which investigates real-world deployment scenarios of multimodal LLMs, and finally, Khaliq et al. [22] specifically applied retrieval-augmented reasoning (ToRAG, CoRAG) to tackle multimodal claims within political context by extracting both textual and image content, retrieving external information, and reasoning subsequent questions to be answered based on prior evidence and achieved a weighted F1-score of 0.85,

surpassing a baseline reasoning method by 0.14 points.

Multilingual Fact-Checking. Misinformation transcends language barriers, necessitating fact-checking capabilities across multiple languages. Zhang et al. [82] in their work conducted a systematic quantitative and qualitative analysis of the multilingual abilities of LLMs by introducing a novel prompt back-translation method. Their result reveals that LLMs excel in transferring learned knowledge across different languages, producing relatively consistent results in translation-equivariant tasks but struggle with translation-variant tasks, requiring careful human evaluation. Therefore, assessing the cross-lingual factual consistency of LLMs should be a primary concern in current research. For instance, Shafayat et al. [83] proposed a pipeline called MultiFAct to evaluate the factuality of long-form multilingual LLM generations. This pipeline demonstrates the feasibility of adapting FActScore with non-English resources as a knowledge source, highlights their potential and limitations, and investigates cultural and geographic biases in LLMs. Quelle et al. [48] found that fact-checking accuracy varied across languages, with translated English prompts often achieving higher accuracy than original non-English ones in terms of multilingual fact-checking, despite claims involving non-English sources. Additionally, due to skewed training data and non-standardized fact-checks, LLMs perform better with English-translated prompts, revealing language bias in multilingual fact verification. We also found a unique and noteworthy work by Siino et al. [84] which introduced a comparative evaluation of state-of-the-art models, including LLMs for detecting fake news. Utilizing the multilingual Profiling Fake News Spreaders on Twitter (PFNSoT) dataset, which includes English and Spanish tweets, they showed that large, pre-trained Transformer models such as RoBERTa, DistilBERT, BERT, XLNet, ELECTRA, and Longformer are not necessarily the optimal solution for the classification task; instead, a convolutional neural network (CNN) is enough to outperform those. Their claim was validated by their proposed CNN model, demonstrating a binary accuracy of 71.5% for the English dataset and 81.5% for the Spanish dataset, where the highest performing transformer models, RoBERTa, only achieved an accuracy of 69.5% in English and ELECTRA achieved 76% in Spanish. Notably, in recent years, Shcharbakova et al. [85] demonstrated a comprehensive benchmark of language models for multilingual fact verification. Their experiments showed that the smaller encoder-based language model XLM-R significantly outperformed a much larger decoder-only LLM, achieving a macro-F1 score of 57.7%, which represents an improvement of approximately 15.8% over the previous state-of-the-art results. They suggested specialized SLMs can be more effective than general-purpose LLMs but also mentioned systematic difficulties in utilizing evidence effectively, alongside notable biases toward frequent categories within imbalanced data settings. In multilingual fact-checking, we

found a distinctive and noteworthy work, performed by Jannah et al. [86], where they assessed the symptom detection capabilities of LLMs in social media posts across seven languages. Their experiment on zero-shot multi-label symptom classification includes large- and small-parameter models from three leading LLM providers: OpenAI, Google Gemini, and Mistral AI. They showed significant performance disparities, with LLMs performing better in European languages than in under-resourced Asian languages, and tending to over-predict specific respiratory illnesses like influenza.

Domain-Specific Fact-Checking. Domain-specific fact-checking is a crucial research area, as the nuances of verifying factual claims can significantly differ across specialized fields like medicine, politics, and climate science, necessitating tailored LLMs and verification systems. While general-purpose fine-tuned LLMs dominate broad tasks, specialized models fine-tuned on specific domains often outperform general-purpose models in those areas [28]. Moreover, proprietary models like Factcheck-GPT are often designed for general-purpose use and not viable in domains like medicine due to restrictions on private data use and lack of fine-tuning [30].

In medical contexts, dedicated efforts include [28], introduced Clinifact and when evaluating LLMs against that, BioBERT achieved 80.2% accuracy, outperforming generative counterparts, such as Llama3-70B's 53.6%, with statistical significance ($p < 0.001$), developing explainable reasoning systems [38], and detecting and correcting hallucinations by integrating multi-source evidence [49]. Similarly, Tans et al. [30] proposed LEAF, which is tailored towards the medical domain as it uses the MedRAG corpus.

Fact-checking is also investigated across various domains, including travel, climate, news information, and claim matching. For example, in the travel domain, Jing et al. [50] used four industry datasets containing chats, reviews, and property information for fact-checking. Political claim verification is explored through studies assessing LLMs' reliability [54] and multimodal retrieval-augmented reasoning systems [22]. In climate science, specialized LLM-based tools address the complexity of climate-related claims [47]. News claims and general misinformation are broadly examined through hierarchical prompting methods in [21], [12]. Other studies investigated the potential of LLMs for fake news detection using datasets such as Weibo21 [91] and GossipCop [92]. Research on news-article veracity has focused on the FA-KES Syrian War corpus [93] and the EUvsDisinfo pro-Kremlin corpus [94], while some platforms enable the development of customized fact-checking systems [44]. [31] trained their SNIFFER model on news domain using the NewsCLIPpings [73] dataset. Additionally, claim-matching methods utilizing LLMs for fact-checking [41, 47, 88] and verification approaches for complex claims [37] further underscore the depth and breadth of ongoing domain-specific fact-checking research.

Enhancing Explainability and Trust. Beyond mere accuracy, the ability of LLM-based fact-checking systems to provide explanations and foster trust is increasingly recognized as vital. Citations and provenance play a central role in this, with [56] highlighting the importance of referencing sources to build user confidence, while [39] emphasizes the need to trace the origin of generated content. In the framework proposed by Quelle et al. [48], agents explain their reasoning and cite the relevant sources. Zhao et al. [14] enhanced fact-checking explainability by utilizing instruction-based LLMs and specialized agents to generate structured reasoning and justifications for sub-claims. Through CLIP-based similarity, ROUGE scores, response ratio analysis, and human evaluation, SNIFFER [31] demonstrated high accuracy and strong persuasive ability in explaining and detecting out-of-context misinformation, and transparency in medical contexts [38]. Credibility assessment frameworks offer structured approaches to evaluate trustworthiness [42]. Additionally, Ghosh et al. [17] investigated the logical soundness of model outputs, a key indicator of reliability. In another study, Leite et al. [55] simplified fact-checking by guiding LLMs to predict individual veracity signals, minimizing hallucinations, and enabling human reviewers to audit and control outputs, which enhanced transparency. Giarelis et al. [18] stated that their LLM-KG framework improves transparency through factual context provided by KGs.

Hierarchical Prompting and Multi-Step Reasoning. A key innovation involves prompting LLMs in a way that structures their reasoning process by decomposing complex fact-checking tasks into smaller, verifiable steps. The underlying idea is that human fact-checkers often break down a claim into sub-claims or evidence checks. LLMs can be prompted to emulate the process, reducing the risk of a single misstep leading to a hallucinated conclusion. For instance, HiSS prompting directs the model to first separate a claim into several subclaims, then verify each subclaim one by one, before finalizing an overall verdict [21]. By forcing the model to focus on one piece of information at a time (in a chain-of-thought style), HiSS achieved superior fact verification performance and reduced hallucination on news datasets, even outperforming fully-supervised baselines. Advance prompting like Hiss outperforms standard CoT Table 7. This demonstrates that well-designed prompting can guide the model to reason more carefully and factually reduce hallucination. In a study, Khalil et al. [22] introduced Chain-of-RAG (sequential) and Tree-of-RAG (branch-and-eliminate hierarchy), which embodied multi-step and hierarchical reasoning. Another example is the PACAR framework [14], which combined LLM-driven planning with customized action reasoning for claims. PACAR consists of multiple modules (a claim decomposer, a planner, an executor, and a verifier) that allow LLMs to plan a sequence of actions, such as performing a web search or a numerical

calculation, and then verify the claim based on collected evidence. Using hierarchical prompting and a multi-step approach, which includes specialized skills like numerical reasoning and entity disambiguation, PACAR significantly outperformed baseline fact-checkers across three different domain datasets.

Overall, research to date illustrates the use of several comprehensive, multi-dimensional approaches to mitigating LLMs’ hallucinations in fact-checking, with notable advances in RAG domain-specific fine-tuning and hybrid methodologies. Nonetheless, guaranteeing robust factual reliability across varied, complex, and dynamically evolving information scenarios continues to pose a major challenge. Future research is expected to prioritize more sophisticated hybrid systems, refined self-correction mechanisms, and more effective human-AI collaboration to strengthen the fact-checking processes [33, 95].

4.3 Datasets for Training and Evaluating Fact-Checking Systems (RQ3)

In this section, we discuss the wide range of datasets used in the training, evaluation, and benchmarking of fact-checking systems, particularly within RAG frameworks and hallucination mitigation strategies. These datasets support various steps in each method, such as claim verification, evidence retrieval, multi-hop reasoning, and hallucination detection. The following accounts for the datasets and their uses in this domain:

Benchmark Datasets for RAG-Based Fact Verification. Core claim verification datasets such as FEVER [96], FEVEROUS [9], and HOVER [97] are widely used to evaluate RAG pipelines, where the model is used to retrieve relevant evidence from Wikipedia or structured sources and then generate a verdict. These datasets provide gold-standard evidence, making them ideal for training retrievers and verifying the accuracy of generated outputs. LIAR [98] and RAWFC [99] further allow the assessment of RAG-based models on political and news-based claims with distinctive complexity and source structures.

Domain-specific Datasets. In domain-specific applications, datasets such as SciFact [100], COVID-Fact [101], MedMCQA [102], BioASQ [103], and PubMedQA [104] are frequently employed in RAG frameworks that are aligned with the biomedical domain. These allow models to retrieve evidence from the medical literature (e.g., via MedRAG) and cross-validate LLM outputs. Given the importance of factual accuracy in these domains, these datasets also serve as valuable benchmarks for evaluating and refining hallucination reduction techniques in sensitive contexts.

Multimodal Datasets. Multimodal datasets such as Multimodal-FEVER (Factify) [105], Post-4V [57], NewsCLIPings [73], and MOCHEG [106] allow extended

fact-checking to vision-language settings. These are especially relevant for evaluating multimodal RAG systems that allow the use of visual and textual information to assess the veracity of claims. Mismatches between image-caption pairs in these datasets test the models’ ability to detect hallucinated or manipulated content across modalities.

Hallucination Detection-specific datasets. To evaluate hallucination detection and correction, specialized datasets such as HaluEval [107], ReaLMistake [108], TruthfulQA [5], and FoolMeTwice [109] provide annotated examples of hallucinated outputs with detailed rationales. These are critical for assessing token-level uncertainty, logical consistency, and explanation alignment in LLMs.

Composite Datasets. Composite benchmarks like Fact-Bench [110], OpenFactCheck [44], and FIRE [111] aggregate multiple datasets (e.g., FacTool-QA, FELM-WK, Factcheck-Bench) to provide diverse evaluation techniques for both retrieval and generation stages. These are specifically valuable for end-to-end RAG evaluation, as they test factuality, consistency, and explainability across all claim types and evidence formats.

Synthetic and Multilingual Datasets. Synthetic and weak supervision datasets such as ClaimMatch [112], LLM-AGGREFACT [10], and CheckThat [113] allow for scalable training and evaluation in low-resource settings. These datasets are often used to pre-train or fine-tune retrievers and scorers within RAG systems, or to assess robustness against adversarial claims and misinformation edits. Additionally, multilingual datasets and some databases, like X-Fact [114], Data Commons Multilingual, and FactStore, help build fact-checking systems that work across different languages. They test whether these systems can find the correct information and give accurate answers, even in non-English settings. This helps make fact-checking tools more significant in global use. Figure 9 visualizes an overview of major dataset types and their domains.

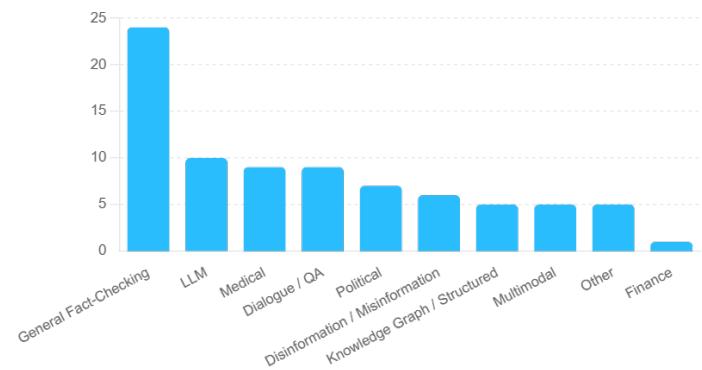


Figure 9: Illustration of major dataset types and domains. The bar chart shows the distribution of datasets across different types and application domains, highlighting their relative prevalence and focus areas.

To provide a clear and comprehensive overview, Table 5 lists 72 key datasets used in fact-checking research. It shows which datasets were used for fact-checking by using RAG, for hallucination reduction, and their type. Out of 72 datasets, 63 use RAG, and 48 are used for hallucination reduction. The use cases of each dataset used for fact-checking are grouped into several key types. To test interactive evidence retrieval and conversational-based fact-checking in multi-turn RAG pipelines, dialogue datasets are used [54, 63]. QA datasets test a system’s ability to accurately answer questions by reasoning through open-domain and multi-hop scenarios, where it retrieves and connects relevant information [62, 97]. Fact-checking datasets provide labelled claims with the most reliable evidence for training and benchmarking claim verification [9, 112]. Complex fact-checking in biomedical contexts is typically facilitated by medical and domain-specific datasets (e.g., SciFact, PubMedQA) [63, 104]. LLM and hallucination datasets capture ungrounded outputs to assess truthfulness and hallucination mitigation. Structured data-to-text datasets can produce more accurate texts from tables or knowledge graphs [8]. Summarisation datasets evaluate the production of concise and reliable text [3]. Misinformation datasets help to find and disprove false or manipulated claims [92, 113], and multimodal datasets check that text–image evidence pairs are consistent [43, 90].

Table 6 represents the key limitations of each dataset type used in fact-checking with LLM. Most datasets are narrow in scope or biased, which affects the reliability of fact-checking in specific domains and languages.

4.4 Prompt Design, Fine-Tuning, and Domain-Specific Training (RQ4)

Prompt design strategies significantly impact the ability of LLMs to perform fact-checking and mitigate hallucination. The choice of strategy influences how the model processes information, accesses knowledge, and generates responses, directly affecting accuracy and performance. The sources explore several key strategies, often contrasting methods that rely solely on the model’s internal knowledge with those that integrate external information retrieval. A visual summary of approaches in prompt design, fine-tuning, and domain-specific training is shown in Figure 10.

4.4.1 Basic Prompting Strategies

Basic prompt strategies primarily rely on internal knowledge and involve presenting the claim or task to LLMs with minimal or no external context beyond the prompt itself. Their effectiveness is heavily dependent on the pre-trained knowledge of the model, which can be a significant limitation due to the potential for hallucination and outdated information [12, 21, 30].

Zero-shot prompting involves providing the LLMs with only the task description and the input claim, without any specific examples. This can include asking the model to directly predict the veracity label of a claim [12, 71]. Frameworks such as FACT-AUDIT [45] used zero-shot inference to evaluate the fact-checking capacity of various LLMs. They gave the lowest average accuracy scores compared to other methods, particularly when relying solely on the model’s internal knowledge. While a simple combination of self-consistency and zero-shot prompt was found to be the most effective overall strategy in a multilingual fact-checking study, this effectiveness was strongly tied to the self-consistency decoding strategy, not necessarily the zero-shot nature itself [32]. On the other hand, frameworks such as PACAR [14], which employed explicit claim decomposition and a dynamic planning mechanism, achieved strong performance in zero-shot settings and outperformed other LLM-based approaches, including few-shot and fine-tuned methods. However, relying solely on internal knowledge for fact-checking is considered unreliable and insufficient, as LLMs are prone to hallucination [12, 21, 30]. Zero-shot prompting without external access means that the model must rely on potentially inaccurate or outdated information stored during training [48]. The analysis of rationales generated through zero-shot CoT indicates unreliability for factuality analysis based on internal memorization [35]. The most significant mitigation discussed is the incorporation of external knowledge retrieval [15, 20, 21, 30, 48, 53].

The Few-Shot Prompting or ICL method involves providing the model with a limited number of examples of the task before presenting the claim to be verified [12, 14]. It utilizes the LLM’s ability to learn from examples provided directly in the prompt ("in-context learning") [20]. Few-shot demonstrations are used in methods like HiSS [21] and BiDeV [37] to guide the LLM through multi-step processes. Few-shot-CoT includes example pairs to guide the reasoning process [47, 71]. A similar approach was used to address the challenge of identifying hallucinations in natural language generation by using the pre-trained Llama 2-7B model in a zero-shot setting. Instead of fine-tuning the model, the authors used specific prompts that included the context and hypothesis sentences, which were provided to the model, to determine whether the context supported the hypothesis with a simple yes or no. The model’s answers were reprocessed to classify them and determine hallucinations or unsupported outputs. They evaluated its performance on two tracks: the model-aware method, with access to specific model checkpoints, and the model-agnostic method, without such access. To keep the process efficient, the method focused on minimal engineering and prevented additional fine-tuning or preprocessing. They claim that the model’s ability to reliably identify hallucinations could be improved even further with adjustments such as fine-tuning it for a specific domain or incorporating more data [115]. In another

Table 5: Datasets and their use in RAG and hallucinations (Halluc.) reduction. It indicates the type of dataset and the tasks for which it has been evaluated.

| Dataset | RAG | Halluc. | Type | Dataset | RAG | Halluc. | Type |
|---------------------|-----|---------|-----------------|-------------------|-----|---------|-----------------|
| Doc2Dial | ✓ | ✓ | Dialogue | LIAR | ✓ | ✗ | Political |
| Topical-Chat | ✓ | ✓ | Dialogue | RAWFC | ✓ | ✗ | Political |
| QReCC | ✓ | ✓ | Dialogue | HALLU | ✓ | ✓ | QA |
| Wizard of Wikipedia | ✓ | ✓ | Dialogue | BioASQ-Factoid | ✓ | ✓ | QA |
| CMU-DoG | ✓ | ✓ | Dialogue | Q2 | ✓ | ✓ | QA |
| FEVER | ✓ | ✗ | Fact-Checking | ASSERT | ✓ | ✓ | QA |
| SciFact | ✓ | ✗ | Fact-Checking | SQuAD | ✓ | ✓ | QA |
| COVID-Fact | ✓ | ✗ | Fact-Checking | CoCoGen | ✗ | ✓ | QA |
| Factify | ✓ | ✗ | Fact-Checking | TriviaQA | ✓ | ✓ | QA |
| AVERITEC | ✓ | ✗ | Fact-Checking | HotpotQA | ✓ | ✓ | QA |
| TruthBench | ✓ | ✓ | Fact-Checking | Natural Questions | ✓ | ✓ | QA |
| LLM-AGGREFACT | ✓ | ✗ | LLM | ELI5 | ✓ | ✓ | QA |
| TruthfulQA | ✓ | ✓ | LLM | NarrativeQA | ✓ | ✓ | QA |
| HaluEval | ✓ | ✓ | LLM | NewsQA | ✓ | ✓ | QA |
| BioASQ-Y/N | ✓ | ✗ | Medical | DROP | ✓ | ✓ | QA |
| MedMCQA | ✓ | ✗ | Medical | FactMix | ✓ | ✓ | QA |
| USMLE | ✓ | ✗ | Medical | DuoRC | ✓ | ✓ | QA |
| MMLU-Medical | ✓ | ✗ | Medical | QuAC | ✓ | ✓ | QA |
| PubMedQA | ✓ | ✗ | Medical | FactScore Dataset | ✓ | ✓ | QA |
| CliniFact | ✗ | ✗ | Medical | Data Commons | ✓ | ✗ | Structured |
| MedQuAD | ✓ | ✓ | Medical QA | WikiBio | ✓ | ✓ | Structured Data |
| MedInfo QA | ✓ | ✓ | Medical QA | ToTTo | ✓ | ✓ | Structured Data |
| LiveQA-Medical | ✓ | ✓ | Medical QA | WebNLG | ✓ | ✓ | Structured Data |
| MEDIQA-RQE | ✓ | ✓ | Medical QA | DART | ✓ | ✓ | Structured Data |
| MeQSum | ✓ | ✓ | Medical Summary | LogicNLG | ✓ | ✓ | Structured Data |
| FakeCovid | ✗ | ✗ | Misinformation | E2E NLG | ✓ | ✓ | Structured Data |
| Multimodal FEVER | ✓ | ✗ | Multimodal | SAMSum | ✓ | ✓ | Summarization |
| COCO | ✗ | ✗ | Multimodal | FIED | ✓ | ✓ | Summarization |
| Objaverse | ✗ | ✗ | Multimodal | XSum | ✓ | ✓ | Summarization |
| Visual Aptitude | ✗ | ✗ | Multimodal | CNN/Daily Mail | ✓ | ✓ | Summarization |
| ADE20K | ✗ | ✗ | Multimodal | Gigaword | ✓ | ✓ | Summarization |
| NewsCLIPpings | ✓ | ✗ | Multimodal | Multi-News | ✓ | ✓ | Summarization |
| Polyjuice | ✗ | ✓ | NLI | Newsroom | ✓ | ✓ | Summarization |
| FactualNLI | ✓ | ✓ | NLI | BigPatent | ✓ | ✓ | Summarization |
| Multilingual FC | ✗ | ✗ | Other | WikiHow | ✓ | ✓ | Summarization |
| PolitiFact | ✓ | ✗ | Political | Reddit TIFU | ✓ | ✓ | Summarization |

Table 6: Identified Limitations of Datasets Commonly Used in LLM-Based Fact-Checking Across Different Domains

| Dataset Type | Key Limitations |
|----------------------------|--|
| Dialogue | Low domain diversity and simple claim structures with weak coverage of real misinformation. |
| Fact-Checking | Biased toward Wikipedia claims with poor temporal and geographic variety and slow tracking of emerging misinformation. |
| QA | Encourages benchmark overfitting and lacks adversarial and multi-hop claims. |
| Medical | Expensive annotation with narrow coverage and poor multilingual and low-resource support. |
| Misinformation / Political | Small size and region specific, with rare updates for new narratives. |
| Structured Data | Clean and synthetic, but fails to capture noisy and conflicting real-world data. |
| Multimodal | Sparse and noisy, with weak standards and limited cross-modal reasoning tests. |
| LLM | Mostly synthetic with short and simple prompts and weak domain-specific coverage. |
| Summarization | Subjective references that reward lexical overlap over factual grounding. |
| Multilingual | Uneven language coverage with high-resource bias and weak cultural generalization. |

instance, an equivalent method is used in a prompt-based approach to few-shot learning to identify hallucinations in Swedish and English paraphrased texts using the Mistral 7B LLM. Task-specific prompts are carefully designed to include instructions and a limited number of reference examples from the training data. The model is given these prompts, the source sentence, and two hypotheses. Then, it is asked to determine which hypothesis contains hallucinated content. Without any additional fine-tuning, the method makes effective use of the model’s reasoning abilities and strategically incorporates representative samples into the prompt to help the model make decisions that will successfully detect hallucinations [116].

ICL has been demonstrated to enhance the performance

of open-source multimodal LLMs (MLLMs) in misinformation detection, occasionally yielding greater improvements than those achieved through prompt ensemble methods [57]. Combining ICL with RAG has been shown to improve accuracy in fact verification [53]. However, some sophisticated few-shot ICL methods like Standard Prompting, Vanilla CoT, and ReAct were surpassed by the HiSS method in news claim verification, highlighting the importance of the specific method prompted [48]. While ICL improves performance, open-source MLLMs using ICL still significantly lag behind state-of-the-art proprietary models like GPT-4V. The effectiveness can vary between models and datasets [57]. The manual prompt design for few-shot examples can be

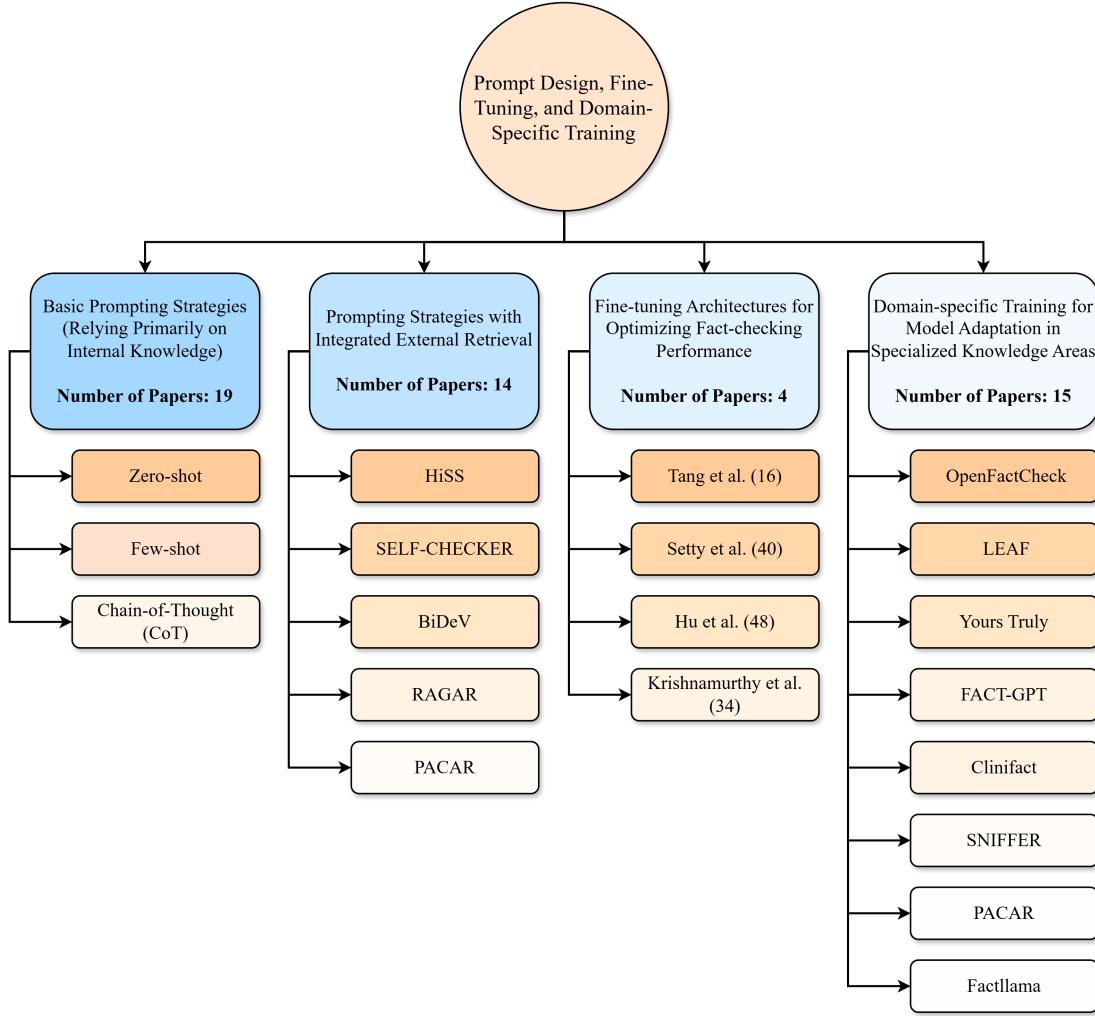


Figure 10: Breakdown of approaches in prompt design, fine-tuning, and domain-specific training for fact-checking with LLMs, categorized into four groups. Each group highlights the number of papers reviewed and the representative techniques.

heuristic [20]. However, providing examples does not guarantee overcoming the dependence on internal knowledge if external information is not integrated [21]. Integrating ICL with RAG or structured step-by-step prompting frameworks is a key mitigation [21, 53].

Other methods, such as CoT prompting, instruct the LLM to output a sequence of intermediate reasoning steps before arriving at the final answer or the veracity label [12, 21]. Zero-shot CoT prompts often include eliciting sentences like "Let us think step by step" [12, 16, 32, 47]. This encourages explicit reasoning. Variants of CoT prompting include English CoT (EN-CoT) [32], which focuses on monolingual reasoning, and CoTVP (CoT Veracity Prediction) [22], which evaluates the truthfulness of reasoning steps. However, studies show that techniques like CoT, designed to improve reasoning, do not necessarily improve the fact-checking abilities of LLMs. They can even have minimal or negative effects on success rates [32]. Models prompted

with CoT may tend to align with the input text rather than verifying its factualness, especially for complex paragraphs, when relying solely on pre-trained knowledge [20]. Vanilla CoT suffers substantially from the issues of fact hallucination and omission of necessary thoughts in the reasoning process [21]. Using the LLM for factuality analysis based on its internal memorization, even with zero-shot CoT, indicates unreliability, likely caused by hallucination [12].

The internal mechanism of LLMs to integrate rationales from various perspectives via CoT can be ineffective for fake news detection [12]. Integrating CoT with external knowledge retrieval, as in Search-Augmented CoT or ReAct [21], is a key approach to mitigate hallucination and thought omission [21]. Furthermore, frameworks that explicitly guide the decomposition and reasoning process, such as FactAgent, are seen as superior to CoT, which primarily acts as a prompting technique [71].

When relying solely on internal knowledge, LLMs

prompted with zero-shot, few-shot, or vanilla CoT struggle in fact-checking complex claims and exhibit hallucination [12, 20, 21, 30]. However, accuracy can be low, and improvements in reasoning via CoT alone do not guarantee better fact-checking performance [32]. Techniques for forcing binary "true" or "false" judgments also do not enhance overall accuracy [72].

4.4.2 Prompting Strategies with Integrated External Retrieval

Several strategies combine prompting with the ability to access and utilize external information sources (e.g., search engines or curated databases) to ground responses and improve factual accuracy. This is a critical aspect for robust fact-checking and hallucination reduction [21, 30, 48, 53].

A variant of CoT that interleaves reasoning traces with task-specific actions, like querying Google Search or the Wikipedia API, allowing the LLM agent to decide whether to search or continue reasoning based on environmental observations [20, 21, 22, 48]. For instance, by accessing external knowledge, ReAct effectively mitigates hallucination failures compared to vanilla CoT and justifies its reasoning with retrieved citations, enhancing verifiability and explainability. Its performance is highly sensitive to the quality and relevance of search results, and relying solely on internal knowledge when external search fails remains a key limitation [21]. Combining ReAct’s action capabilities with more structured decomposition and step-by-step verification methods, such as HiSS or SELF-CHECKER, can address thought omissions and improve overall performance [20, 21].

Search-Augmented CoT augments vanilla CoT by using the original claim as a search query to retrieve background information, which the LLM incorporates into its thought chain. This approach improves over vanilla CoT by utilizing external knowledge, but can fall short of methods like Standard Prompting or HiSS, which indicates that querying solely with the claim may output insufficiently detailed results. To mitigate this, more sophisticated query generation strategies and integration methods are needed [21]. HiSS is a few-shot method that prompts the LLM to perform claim verification in fine-grained steps by decomposing claims into subclaims and verifying each step-by-step, raising questions and optionally using web search when confidence is low. It significantly surpasses few-shot ICL counterparts like Standard Prompting, Vanilla CoT, and ReAct in average F1-score, offering superior explainability through enhanced coverage and readability while substantially reducing hallucination and thought omission [21]. However, HiSS still struggles with integrating updated information from mixed sources and incurs high computational costs due to multiple LLM calls, with its performance remaining sensitive to prompt design [20].

Frameworks like SELF-CHECKER [20], BiDeV [37], RAGAR [22], and PACAR [14] integrate prompting and

RAG by decomposing fact-checking into subtasks (e.g., claim detection, retrieval, sentence selection, verdict prediction) and using prompts often with few-shot examples to generate search queries, select evidence, and perform step-by-step verification while explicitly incorporating retrieved documents. Incorporating external knowledge via RAG significantly boosts accuracy over internal-only approaches: SELF-CHECKER [20], BiDeV [37], RAGAR [22], and PACAR [14] all show substantial gains, with specialized modules such as the Claim Atomizer and Fact-Check-Then-RAG further enhancing predictive power and explanation quality.

RAG-based methods can be hampered by overwhelming context windows [48], outdated or variable search results [20, 22], high computational costs, prompt sensitivity, and manual prompt design [20], and they may still miss fine-grained details even when grounded [30]. To mitigate these issues, current strategies include using IR functions (e.g., BM25) to distill relevant content [30, 48], using multi-agent [37, 75], explicit decomposition and filtering [14, 37, 42], feedback loops [51], and refined RAG processes [30, 42]. Reliance solely on an LLM’s internal knowledge, whether via zero-shot, few-shot, or vanilla CoT prompting, is unreliable for fact-checking and prone to substantial hallucination [12, 21, 30], as zero-shot CoT and vanilla CoT often succumb to memorization pitfalls [12, 16]. Mitigating hallucination primarily involves incorporating external knowledge through ReAct, Search-Augmented CoT, HiSS, and other RAG frameworks [21, 30, 53].

In conclusion, while basic prompting strategies like zero-shot, few-shot, and vanilla CoT offer foundational ways to interact with LLMs for fact-checking, their accuracy and reliability are severely limited by the reliance on potentially flawed internal knowledge [12, 21, 30]. The most effective approaches, as highlighted by the sources, involve prompting strategies that explicitly integrate external knowledge retrieval through frameworks such as ReAct, HiSS, SELF-CHECKER, BiDeV, RAGAR, and PACAR [20, 21, 37, 48, 53]. These methods use prompts to guide LLM through processes that involve external data access, significantly improving accuracy and mitigating hallucination by grounding the model’s responses in evidence [20, 21, 53].

Prompting is also used to generate explanations, although the utility and reliability of these explanations can vary [21, 33]. Limitations across strategies include sensitivity to prompt wording, computational cost, and the need for more robust and automated design methods [20].

4.4.3 Fine-tuning Architectures for Optimizing Fact-checking Performance

Fact-checking performance is primarily optimized through two primary approaches: (1) the development and fine-tuning of specific model architectures and (2) the design of advanced prompting strategies for LLMs. These methods

are often combined within complex fact-checking pipelines.

Fine-tuning smaller transformer models on synthetic data. This approach fine-tunes pre-trained transformer models on structured synthetic data, often combined with standard entailment datasets, to teach them nuanced fact-checking against grounding documents [10]. The strategy focuses on generating challenging training instances that help models verify atomic facts across multiple sentences, with models such as MiniCheck-FT5, RBTA, and DBTA outperforming larger LLMs such as GPT-4 in specific benchmarks like LLM-AGGREFACT [10]. Notably, MiniCheck-FT5 achieves a 4.3% improvement over AlignScore using a significantly smaller dataset. Difficulty aggregating evidence and reasoning over multiple facts, the method mitigates these through targeted synthetic data and simple aggregation strategies like majority voting [11].

Fine-tuning SLMs for task-specific performance. This method focuses on fine-tuning SLMs for task-specific applications like fake news detection. The strategy includes training SLM directly on the target dataset and using LLM-generated rationales via architectures like the ARG network, with a distilled version (ARG-D) for efficiency [12]. Fine-tuned BERT models have outperformed GPT-3.5 in fake news detection, and ARG/ARG-D exceed baseline methods that combine both SLM and LLM capabilities [12]. LLMs' difficulty in fully utilizing their reasoning for domain-specific tasks and their inability to fully replace SLMs in these contexts. Mitigation strategies include employing LLMs as rationale providers to guide SLMs, as well as exploring advanced prompting techniques and model combinations to achieve improved performance [12].

Instruction Fine-tuned LLMs as Verifiers within a Pipeline Framework. This framework integrates an instruction-fine-tuned LLM as a verifier within a fact-checking pipeline, where the LLM evaluates claims based on contextually retrieved evidence. The process includes claim atomization using Mistral-7B, relevance-based evidence retrieval and re-ranking, and final inference by the LLM, producing interpretable credibility reports [42]. The "Yours Truly" framework achieves a 94% F1-score, significantly outperforming other systems, with the Claim Atomizer alone boosting performance from 64% to 93% [42].

4.4.4 Domain-specific Training for Model Adaptation in Specialized Knowledge Areas

Domain-specific adaptation and fine-tuning are highlighted as crucial strategies for enhancing the performance and reliability of models, particularly LLMs, in automated fact-checking within specialized knowledge areas. The need for such domain specificity arises from the observation that the factual accuracy and vulnerability of LLM to hallucinations can vary significantly between different domains. While

general-purpose models may perform well in broad areas, models fine-tuned or adapted for specific domains, such as medicine or science, often demonstrate superior performance in those particular fields. Sources discuss the application of fact-checking techniques across various specialized domains, including medical/biomedical, political, scientific, law, general biographic, and news [14, 28, 30, 31, 44].

Several approaches are explored to achieve domain adaptation in fact-checking systems. One involves fine-tuning smaller transformer models on target datasets specific to a domain or task within fact-checking, such as claim detection or veracity prediction, which has shown surprising efficacy and can outperform larger, general LLMs in specific contexts [11, 48]. Another approach directly involves fine-tuning LLMs themselves or using techniques like instruction-tuning on domain-relevant data or tasks [30, 31, 40, 42, 46, 48]. Frameworks like OpenFactCheck are proposed to allow users to customize fact-checkers for specific requirements, including domain specialization [44].

4.4.5 Comparative Summary and Trends

From straightforward prompting to more complex fine-tuning and domain-specific training, the methods for using LLMs in fact-checking are changing. To inform general-purpose models such as GPT-3.5 and GPT-4, initial research focused on zero-shot and few-shot in-context learning. However, in fact-checking benchmarks, a notable trend indicates that smaller, optimized models can surprisingly beat larger LLMs, providing a more economical and effective solution [11].

Prompt engineering has also progressed beyond simple CoT. The rise of organized hierarchical prompting techniques, such as HiSS, which break down complicated claims into verifiable steps to reduce hallucinations and improve reasoning, is a definite trend [21, 38]. Furthermore, domain-specific adaptation is becoming more and more important. Using specialized datasets and knowledge bases, models are being particularly trained or prompted for difficult domains such as news, climate science, and medicine to increase the accuracy when general knowledge is inadequate [49, 75].

Integrating external domain-specific knowledge through methods like RAG or incorporating Knowledge Graphs is also a significant strategy to augment models with the necessary specialized information, sometimes in conjunction with fine-tuning or adaptation [15, 18, 36, 30, 46]. For example, the LEAF approach enhances medical question answering by integrating fact-checking results into RAG and using fact-checks for supervised fine-tuning [30]. Evaluation datasets tailored to specific domains or types of claims, such as SciFact for scientific claims, Climate-FEVER for climate science, EXPERTQA spanning multiple fields, or medical QA datasets, are utilized to benchmark the effectiveness of these domain-adapted systems [10, 14, 28, 30, 75].

While interventions like fine-tuning and claim normalization improve robustness, performance can still degrade when evaluated significantly out-of-domain across different topics or platforms [40]. The sources collectively underscore that effective fact-checking in specialized areas often requires models or frameworks specifically tailored to the domain’s knowledge and nuances, moving beyond one-size-fits-all general approaches [28, 30, 40, 44].

4.5 Integration of RAG in Fact-Checking (RQ5)

RAG is a hybrid framework designed to enhance LLMs by integrating the retrieval of external knowledge into their generation process [3, 56]. This approach is pivotal for grounding LLM outputs in external evidence, thereby mitigating common issues such as hallucination, the generation of factually incorrect or nonsensical information, and reliance on potentially outdated internal knowledge [3, 15, 33]. Unlike methods solely based on fine-tuning the model’s internal parameters, RAG employs external data sources, such as web search results or curated knowledge bases, to inform the LLM’s responses during inference [15, 56]. A significant advantage of RAG is its ability to produce responses that are factually accurate and reliable, and offer improved transparency and credibility by providing explicit citations to the **external sources used** [48, 56, 117]. This capability allows models to access and utilize current information, overcoming the limitations of knowledge cutoffs inherent in their training data [22]. An overview of the workflow of a basic RAG-based system is presented in Figure 11.

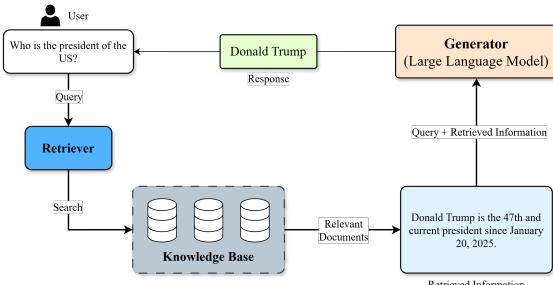


Figure 11: Workflow of a RAG system for factual question answering.

In the domain of fact-checking, RAG systems play a crucial role in automating and improving the verification process [48]. LLM agents empowered with RAG can phrase search queries based on claims, retrieve relevant external contextual data, and utilize this information to assess the veracity of statements [48, 56]. This integration of external knowledge through RAG leads to enhanced accuracy in fact-checking, enabling the extraction of relevant evidence to support veracity predictions [22, 48, 53]. Approaches like

Fact-Check-Then-RAG utilize the outcomes of fact-checking to refine the retrieval process itself and ensure that retrieved information specifically enhances factual accuracy [30]. RAG also supports more complex scenarios, including multimodal fact-checking, where it is used to extract both textual and image content and retrieve external information for reasoning [22, 57]. Furthermore, RAG-based systems can provide reasoned explanations for their verdicts, improving the interpretability of the fact-checking process [14, 48]. Variations like FFRR utilize fine-grained feedback from the LLM to optimize retrieval policy based on how well documents support factual claims [51]. While some approaches explore reducing reliance on external retrieval by comprising the LLM’s internal knowledge, RAG is generally considered essential for effective fake news detection [57, 71].

Despite its advantages, implementing RAG, particularly in specialized domains, presents several challenges and limitations [3, 76]. A significant hurdle is the requirement for efficient and accurate retrieval of relevant evidence at scale, which can be a computational bottleneck [3]. The effectiveness of RAG is inherently dependent on the assumption that pertinent information is readily available and accessible within the external knowledge sources used, such as search engines [21]. This assumption may not hold for all information, especially in specialized or low-resource domains where relevant knowledge might be obscure, non-digitized, or exist in formats not easily indexed [21]. Retrieving and processing large volumes of external data can also overwhelm the LLM’s context window, necessitating sophisticated techniques for selecting and consolidating the most critical information [15, 48]. Standard RAG-based methods can inadvertently introduce noise or irrelevant information, potentially hindering the LLMs’ performance rather than improving it [30, 51]. In specialized fields like healthcare, general RAG models may struggle due to a lack of the nuanced understanding required for accurate fact-checking, highlighting the need for tailored or potentially fine-tuned approaches or integrating domain-specific knowledge bases [18, 30]. Complex fact-checking tasks, such as interpreting conflicting evidence, handling claims with insufficient external information, or performing causal reasoning with fragmented information across documents, remain challenging for LLMs even with RAG [37, 53, 42]. Furthermore, the dynamic nature of information requires constant updates to external sources, and relying solely on search results can lead to diluted credibility if misinformation is widely reported [71].

For inefficient and computationally expensive evidence retrieval, researchers have developed lightweight frameworks like Provenance [39], which use compact and open-source NLI models for verification instead of LLMs. FIRE [34], a framework that is designed for time and cost efficiency through an iterative process. Reinforcement retrieval models further show that a properly trained retriever does not

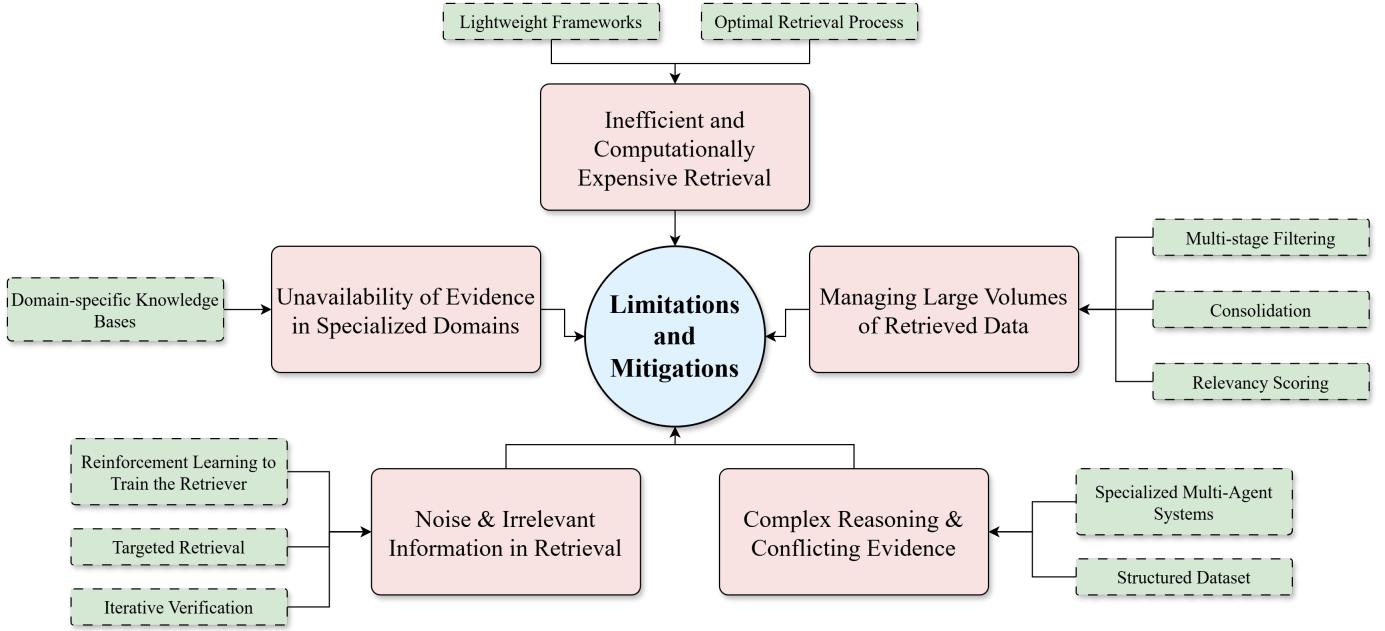


Figure 12: Limitations in RAG-based fact-checking and mitigation Strategies. The red boxes illustrate the limitations, while the green boxes represent the corresponding mitigation strategies.

add significant overhead during inference, as the costly feedback loop is only part of that training phase [51]. Due to the unavailability of evidence in specialized or low-resource domains, frameworks like LEAF [30], designed for the medical domain, use a specialized corpus such as MedRAG, which includes PubMed and textbooks, instead of relying solely on Google Search. The challenge of fact-checking in low-resource languages is mitigated by translating the claims into high-resource languages, such as English, in some literature [48].

To prevent overwhelming the LLM's context window and to efficiently manage large volumes of retrieved information, frameworks such as Provenance [39] employ a Relevancy Score in combination with TopK/TopP selection modules to filter the most critical evidence before it reaches the verifier. Similarly, LLM-AUGMENTER incorporates a "Knowledge Consolidator" that prunes irrelevant data and synthesizes the remaining evidence into concise reasoning chains [51]. Reinforcement Retrieval further demonstrates that limiting the number of documents, typically to the top three or four, can be more effective than including a larger set, which may introduce noise [51]. For mitigating noise and irrelevant information introduced by standard RAG methods, Reinforcement Retrieval addresses this directly using reinforcement learning to train the retriever; feedback from the LLM verifier acts as a reward signal, teaching the retriever to select more useful and factually relevant documents [51]. The LEAF framework uses a "Fact-Check-Then-RAG" approach, where an initial fact-check on the LLM's output is used to guide a more targeted and accurate retrieval process

[30].

Furthermore, to handle complex reasoning, conflicting evidence, and insufficient information, the PACAR framework employs a dynamic planner that can deploy tailored agents for tasks like numerical reasoning and entity disambiguation [14]. The LoCal framework is a multi-agent system specifically designed to handle the complexities of logical and causal fact-checking. For scenarios with conflicting or insufficient evidence [52], the AVERITEC dataset was created with a "Conflicting Evidence" label and a structured question-answer format that can represent evidential disagreements, providing a basis for training models to better navigate such ambiguity [52].

4.5.1 Comparative Summary and Trends

The incorporation of RAG, which has been a key strategy in LLM-based fact-checking, has established a trend away from dependence on static, internal knowledge. A complex claim is typically divided into verifiable sub-claims in the basic RAG pipeline [21, 49], followed by the collection of relevant external evidence, its combination with a verifier model, and a final decision [46, 118]. It is evident that in recent years, this linear pipeline has changed into more intelligent, dynamic, and effective systems. One significant advancement is the shift to agent-based and iterative frameworks. Systems like FIRE [34] and LLM-AUGMENTER [15] can handle complex, multi-hop claims more successfully because they employ repeated cycles of retrieval and verification rather than a single retrieval step. Multi-agent systems like

LoCal and PACAR [14], which employ planning modules to dynamically choose which tools or reasoning processes to employ, further develop this. The growing complexity of the retrieval and verification procedure itself is another significant development. Frameworks such as Provenance [39] employ lightweight models to score and filter evidence for relevance rather than just providing an LLM’s raw search results. As investigated in the Reinforcement Retrieval framework, there is also a shift toward optimizing the retriever for the downstream fact-checking job by employing strategies like reinforcement learning to transmit feedback from the verifier back to the retriever. An overview of the limitations and mitigation strategies is summarized in Figure 12.

5 Discussion

Our review explored the rapid adoption of LLMs in the complex task of automated fact-checking. By examining a wide range of current studies, we have mapped out how we measure their success, the persistent problem of models generating false information (hallucinations), and the essential role and limitations of the datasets they rely on. We also looked into methods for improving LLMs, from prompt engineering and fine-tuning to the increasingly vital use of RAG. The research landscape reveals a field that is buzzing with innovation and showing great promise. It simultaneously highlights significant and complex hurdles that remain. If LLMs are to become truly reliable tools in the global effort against misinformation, these challenges demand ongoing and rigorous investigation [3, 48].

Evaluation metrics. In the evaluation metrics domain (RQ1) for LLM-based fact-checking, we can see the clear transition in classification scores towards more sophisticated, holistic, and context-aware frameworks. The rise of rigorous benchmarks such as LLM-AGGREFACT [10] and the AVERITEC dataset [52], along with new centralized tools such as OpenFactCheck [44], marks a major step towards creating shared ways to measure how accurate LLMs are. However, even with this progress, many important issues remain unresolved. Furthermore, making models tough enough to handle misinformation that adapts, often as an adversary, including claims with subtle edits or those that change over time, remains a major hurdle [40]. The lack of strong methods for clear explanation and adaptation to new types of risks makes it harder for everyday users to judge whether the answers from LLMs are true. This shortfall also weakens the usefulness of these models in fast-changing real-life situations.

Two key challenges are now drawing more attention: teaching models how to check their work and catch their own mistakes [29], and finding reliable ways to measure how

closely their responses match the source material [50]. Despite advances in automated and LLM-driven assessments, Human Evaluation remains essential for nuanced aspects like explanation quality and contextual appropriateness [72], although it is resource-intensive. In general, while progress is evident, significant lacunae persist. There is a pressing need for standardized metrics that robustly assess the quality of LLM-generated explanations [31], the resilience of the model against evolving misinformation [40], and the logical integrity of reasoning pathways [17]. These developments underscore an urgent and ongoing need for metrics that can holistically evaluate not only the veracity of claims, but also the provenance of supporting evidence [39] and the logical integrity of the LLM’s reasoning process.

Hallucination in LLMs. The tendency of LLMs to hallucinate (RQ2), that is, to generate outputs that are linguistically fluent and coherent yet factually misleading or entirely unsubstantiated, remains a significant barrier to their trustworthy deployment in sensitive, high-stakes applications such as fact-checking [3, 21]. RAG has become a foundational mitigation strategy, designed to ground LLM responses in verifiable external knowledge and reduce the models’ reliance on their internal, and potentially flawed or outdated, parametric knowledge [15, 46, 48, 118]. New efforts to boost how truthful language models are include adding systems that automatically give feedback, helping refine answers over multiple tries [15]. Tools like Self-Checker are also being built. These are smart correction modules that let the model review and fix its own mistakes [20]. On top of that, researchers are testing ways to bring together evidence from several sources, with models like MEDICO showing how this kind of fusion can work [49]. This persistence of hallucinations is partly due to the inherent “black box” nature of current LLMs, the difficulty in comprehensively modeling nuanced world knowledge, and the escalating sophistication of adversarial attacks designed to exploit model vulnerabilities [50].

Datasets for Fact-Checking. The datasets (RQ3) utilized for the training, fine-tuning, and rigorous evaluation of fact-checking LLMs are pivotal to their ultimate performance and generalizability. While foundational datasets such as FEVER [96] have been instrumental in catalyzing early research, the field is increasingly recognizing the necessity for more specialized, challenging, and contextually rich benchmarks. Illustrative examples include Clinifact, which is tailored for claims within the domain of clinical research [28], AVERITEC, with its distinct emphasis on real-world claims that necessitate web-based evidence retrieval [52], and BINGCHECK, specifically designed for assessing the factuality of LLM-generated text [20]. The intrinsic characteristics of these datasets, including their composition, scale, annotation quality, topical diversity, and potential inherent

Table 7: Quantitative Synthesis of Key Findings on Fact-Checking Methodologies.

| Claim & Core Insight | Quantitative Evidence & Context | Papers |
|---|--|--------|
| Domain-Tuned SLMs Can Outperform Larger LLMs | <p>A fine-tuned BERT model achieved a +9.0% relative improvement in F1-score over the best-performing GPT-3.5-turbo configuration (76.5% vs. 70.2%) on the GossipCop dataset [12].</p> <p>A supervised ROBERTa-Base classifier outperformed zero-shot GPT-3.5-turbo by +9.6 absolute points in F1-macro on the FA-KES dataset (52.9% vs. 43.3%) [55].</p> <p>A fine-tuned BioBERT for clinical claims achieved 80.2% accuracy on Clinifact, significantly outperforming both zero-shot (34.3%) and fine-tuned (53.6%) Llama3-70B [28].</p> <p>MiniCheck-FT5 model (770M parameters), fine-tuned on specially generated synthetic data, achieved an average Balanced Accuracy of 74.7% on the LLM-AGGREFACT benchmark, which was statistically comparable to GPT-4's score of 75.3 GPT-4's score of 75.3% [10].</p> | 4 |
| Advanced Prompting is Better than Standard CoT | <p>The HiSS (Hierarchical Step-by-Step) method surpassed vanilla CoT by +9.5 absolute points in F1-score on the RAWFC dataset (53.9% vs. 44.4%).</p> <p>On the LIAR dataset, HiSS outperformed vanilla CoT by +7.1 points in F1-score (31.3% vs. 24.2%). HiSS also outperformed the more advanced ReAct agent framework by +4.1 absolute points on RAWFC [21].</p> <p>At a detailed 5-class classification level (Level 2), the CLIMINATOR framework achieved an accuracy of 72.7%. This was substantially better than the GPT-4o advocate alone, which only achieved 56.5% accuracy on the same task[75].</p> | 2 |
| RAG Provides Significant Performance Gains | <p>Providing external context (RAG) to GPT-4 on the PolitiFact dataset increased its accuracy on non-ambiguous verdicts from 75% to 89%. The accuracy on "true" claims jumped by +13.62 absolute points [48].</p> <p>The Fact-Check-Then-RAG method improved Llama 3 70B's accuracy on the PubMedQA dataset from 60.60% to 73.60% (+13.0 absolute points) by using fact-checking results to guide retrieval [30].</p> <p>An RAG pipeline using Mixtral achieved a 0.780 F1-score on the 'Refuted' class on the Averitec development set [53].</p> <p>On the RAWFC dataset, the fine-tuned FactLLaMA model with external knowledge achieved a Macro-F1 score of 0.5565. This significantly outperformed the same model fine-tuned without external knowledge, which only scored 0.5376 [46]</p> <p>On the LIAR-RAW dataset, "Direct Prompting" of the LLM gave an F1 score of 27.0%. Simply adding a frozen (non-optimized) retriever in the FFRR-frozen setup increased the F1 score to 30.1%, demonstrating the inherent benefit of RAG. [51]</p> | 5 |
| Hybrid and Multi-Agent Systems are More Effective | <p>Compared to strong LLM baselines like Flan-T5 and ChatGPT, the LoCal multi-agent system showed an average performance improvement of up to 7.75% in the gold evidence setting and up to 6.17% in the open book setting [16].</p> <p>The hybrid SLM+LLM ARG network improved F1-score over its BERT-only baseline by +3.1 absolute points on the Weibo21 dataset (78.4% vs. 75.3%) [12].</p> <p>The PACAR framework, with specialized agents, outperformed a general ChatGPT baseline by +16.9 absolute points on HOVER 4-hop claims (72.61% vs. 55.72%) [14].</p> <p>The FACT-AUDIT adaptive multi-agent framework demonstrated superior evaluation robustness over static, single-agent pipelines by dynamically assigning roles [45].</p> | 4 |
| Automated Feedback Mechanisms Reduce Hallucinations | <p>The LLM-AUGMENTER system, using a BM25 knowledge consolidator and automated feedback, improved the KF1 score to 37.41 over the GPT KF1 score of 31.33 [15].</p> <p>The Self-Checker framework improved label accuracy on the BINGCHECK dataset from 21.0% (ReAct baseline) to 63.4% [20].</p> <p>Medico's multi-source evidence fusion and correction loop improved hallucination detection F1-score by +34.4 points over its baseline on the HaluEval dataset [49].</p> <p>The Visual Fact Checker uses object detection and VQA models as automated "tools" to verify and correct initial caption proposals, significantly reducing hallucinations in detailed image captions [43].</p> | 4 |
| Fine-tuning on Synthetic Data Boosts Performance | <p>The MiniCheck-FT5 model, trained on generated synthetic data, achieved a Balanced Accuracy of 74.7% on the LLM-AGGREFACT benchmark, a +4.3 absolute point improvement over the previous state-of-the-art. An ablation study showed that removing this synthetic data caused the model's performance to drop by -14.8 absolute points [10].</p> <p>FACT-GPT was trained on a synthetic dataset of contradicting, entailing, or neutral claims generated by GPT-4, which enabled a smaller, specialized LLM to match the claim-matching accuracy of larger models [88].</p> | 2 |

biases (e.g., political, cultural, or temporal) profoundly influence model performance, the ability of models to generalize to unseen domains or claim structures, and, consequently, the perceived effectiveness of various fact-checking methodologies [54, 95]. The development and meticulous curation of robust multilingual datasets [32] also represent a critical frontier to advance the global applicability and equity of LLM-based fact-checking technologies.

Optimization Strategies and Domain-specific Training. Prompt engineering, fine-tuning strategies, and domain-specific training (RQ4) have been shown to significantly modulate the efficacy of LLMs in complex fact-checking tasks. Advanced prompting techniques, such as hierarchical step-by-step verification methods [21] or structured reasoning frameworks like PACAR [14] and BiDeV [37], frequently demonstrate superior outcomes when compared to simpler, more direct prompting approaches. The application of zero-shot and few-shot learning paradigms is also being actively explored for a range of related tasks, including claim matching [41, 47, 88], indicating a strong potential for efficient adaptation of LLMs with limited task-specific data. Furthermore, the surprising efficacy of smaller, specifically fine-tuned transformer models in certain fact-checking contexts [11] compellingly suggests that model scale is not the sole, nor always the primary, determinant of performance. This finding challenges the prevailing "bigger is better" narrative in LLM development and highlights the potential for more resource-efficient, specialized models to achieve competitive or even superior performance in targeted fact-checking applications, particularly when data and computational budgets are constrained. Domain-specific adaptations, which may involve the utilization of LLM-predicted credibility signals [55] or the development of highly specialized systems for critical domains such as medicine [38], are proving essential for achieving the nuanced understanding and reliable fact verification required in these contexts.

The Integration of RAG. Incorporating LLM with RAG (RQ5) is increasingly recognized as a central and indispensable strategy for enhancing the factuality of LLM outputs. This is achieved by providing models with dynamic access to external, often real-time, knowledge sources, thereby augmenting their inherent capabilities [15, 46, 48, 118]. A framework such as FIRE [34] is being developed to optimize iterative retrieval and verification processes, aiming for greater efficiency and accuracy. Other lines of research explore the application of reinforcement learning techniques to refine and optimize retrieval strategies [21]. Nevertheless, significant challenges remain within the RAG pipeline. These include the efficient and precise retrieval of truly relevant evidence from vast, heterogeneous, and often noisy information spaces; the effective fusion of information derived from multiple, potentially contradictory, sources [14]. Extending

RAG to effectively handle multimodal inputs [22, 57] and optimizing the timing and contextual relevance of information recommended by RAG-empowered agents [61] remain active and critical areas of ongoing investigation.

6 Open issues and challenges

While LLMs have advanced fact-checking capabilities, several core challenges remain. These include mismatches between fluency and truth, domain limitations, and weak reasoning integration.

Despite the advancement in fact-checking using LLMs, one common challenge remains the gap between model-generated responses' linguistic quality and factual accuracy. Today's criteria of evaluation are inclined to appreciate language or text overlap with references, which can overlook basic fact errors. It guides to feedbacks that look questionable but sound good and is inaccurate. These responses can receive arbitrarily high scores, which makes the system seem more accurate than it is [48, 58, 59].

Existing models often execute much better on small synthetic datasets, but often fail to generalize well to real-world circumstances. The reason lies in the limited datasets with low complexity, variation in topics, and multilingual texts. Consequently, the models do not handle the variability and nuance of real-world data across languages and topics, which implies the need for more realistic and broad training and test corpora [60, 88]. Additionally, Siino et al. added that transformer-based models generally perform the worst in terms of standard deviation [84]. However, RAG techniques have a stronger evidence-based foundation in LLM, but retrieval is imperfect. Fact-correctness is often weakened by the quality of the retrieved information, noisy or irrelevant documents, and when there is insufficient context available. Similarly, advanced prompting techniques, such as CoT reasoning or multi-agent collaborations, still must be precisely fine-tuned, but they also remain vulnerable to cascading errors in output generation [14, 76, 119].

One of the areas with great potential that is still underexplored lies in the integration of LLMs with symbolic reasoning or structured logic-based systems. These systems can improve interpretability and fact resilience in fact-checking pipelines. But work in this area is still in an early stage, and significant effort is needed to design scalable, proper architectures that can balance the respective strengths of neural and symbolic methods [61, 75]. Table 8 provides an organized summary of the existing issues and challenges discussed in this section.

Table 8: Identified issues in LLM-based fact-checking and their implications. This includes observed behaviors, underlying causes, and the significance of each issue for reliable deployment.

| Issue/Challenge | Observed Behavior | Implication | Why does it matter? | Ref. |
|---|---|---|---|----------------------|
| Mismatch Between Output Quality and Factual Accuracy | Models write very fluent and convincing text. | High-quality language does not mean the facts are correct. | 1. Current evaluation methods favor “sounding good” over “being accurate.” 2. Models may get high scores for responses that look right but contain factual errors. | [48] [58] [59] |
| Limited Relevance Across Domains and Languages | Models perform well on simple, synthetic, or English-only datasets. | Struggle with real-world, complex, or multilingual data. | 1. Fact-checking needs to work across many subjects, topics, and languages. 2. Limited data variety in training/testing leads to poor generalization. | [60] [88] |
| Challenges in Retrieval and Prompting Mechanisms | Use of RAG brings in external evidence | 1. Retrieval is often imperfect (brings irrelevant or noisy info). 2. Advanced prompting (CoT, multi-agent) still leads to error cascades. | 1. Fact-checking relies on reliable evidence retrieval and reasoning chains. 2. Weaknesses here mean incorrect or unsupported conclusions. | [14] [76] |
| Lack of Integration with Symbolic or Structured Reasoning | Current LLMs rely mostly on pattern recognition, not logic. | 1. Little integration with logic/symbolic systems. 2. Models can not follow strict, logical reasoning pipelines. | 1. Symbolic reasoning would make fact-checking more robust and explainable. 2. Lack of it = less trustworthy and harder-to-monitor systems. | [75] [61] |

7 Critical analysis of future research agendas

LLMs show great promise in automating fact-checking, but their use also highlights a range of ongoing problems that still need attention. In the future, research must take a thoughtful and future-focused approach. If these models are to become trustworthy, accurate, and ethically sound tools in fact-checking, the gaps in today’s research, for example, the unreliability of retrieved evidence, the difficulty in verifying complex claims requiring multistep reasoning, and the challenge of mitigating factual hallucinations, need to be tackled head-on. What follows is a breakdown of key areas where further study could really make a difference, each pointing to where progress is most needed.

Evaluation Framework Advancement. A fundamental imperative lies in transcending current evaluation metrics, which often inadequately capture the nuanced performance characteristics of LLMs in fact-checking tasks [3, 44]. Future research must prioritize the development of sophisticated, multi-dimensional frameworks. This includes establishing standardized metrics for explainability and interpretability that demonstrably correlate with human cognitive trust and facilitate diagnostic understanding of model failures [31, 38, 56, 120, 121, 122, 123, 124]. Furthermore, the creation of dynamic evaluation suites (testing systems that can actively change and adapt over time, rather than relying on fixed, unchanging datasets) to rigorously test resilience against evolving misinformation tactics and sophisticated adversarial attacks is paramount [40], moving beyond static benchmarks. Currently, the development of precise metrics for fine-grained faithfulness and verifiable provenance tracking [39, 50] is crucial to ensure that LLM outputs are not merely plausible but demonstrably grounded in credible

evidence.

Factual Hallucination Mitigation. The mitigation of LLM-generated hallucinations requires a strategic shift from reactive correction to proactive prevention mechanisms embedded within LLM architectures and training paradigms [14, 49]. Currently, optimizing RAG systems to effectively navigate complex and noisy information environments [21, 37] and enabling dynamic and reliable knowledge updates within LLMs [3] are critical to ensure accurate and consistent factual grounding. Without these, RAG systems risk amplifying, rather than rectifying, inaccuracies.

Logical Consistency, Reasoning, and Calibrated Trust. The ultimate efficacy of LLMs in fact-checking is critically dependent upon their capacity for robust logical reasoning and their ability to engender warranted user trust. Future endeavors must explore formal verification methods (mathematically and logically ensuring that the model’s reasoning process is sound and its conclusions are consistent) to enhance the logical consistency of LLM outputs [17] and significantly deepen their reasoning capabilities for complex, multihop, and inferential claims [49]. The degree to which users believe the fact-checks offered by LLMs must also be thoroughly investigated. The goal of this study should be to appropriately “calibrate” or modify that trust to a suitable degree. The purpose is to encourage people to critically evaluate the information provided rather than relying too much on these automatic checks [56, 72].

Multimodality and Multilinguality. Given that misinformation transcends unimodal (ie, English) texts, even though some studies have tried to incorporate multimodal and multilingual fact-checking [32, 43, 48, 57], there is still a need for a significant expansion of LLM fact-checking capabilities. The development of robust multimodal systems capable of verifying claims that integrate textual, visual,

Table 9: Identified gaps from the proposed RQs, and future research paths in fact-checking with LLMs.

| RQs | Research Gaps | Potential Research Paths |
|------------|---|---|
| RQ1 | Evaluation and Benchmarking Challenges | <ol style="list-style-type: none"> 1. Develop a new evaluation matrix that not only limits itself to overlapping or semantic scores but also incorporates the factual correctness and the reasoning capabilities of LLMs, along with real-world dynamics and practicality. 2. Establish robust metrics and methodologies for evaluating the human-computer interaction aspects of fact-checking systems in terms of clarity, actionability, and persuasiveness of explanations. 3. Create evaluation frameworks that can assess a fact-checking system's ability to handle temporally sensitive claims, outdated evidence, and the "freshness" of information. |
| RQ2 | Trust and Reliability | <ol style="list-style-type: none"> 1. Develop automatic detection and correction methods for hallucinated LLM outputs, including uncertainty quantification. 2. Optimize retrieval strategies with reinforcement learning and iterative verification to improve efficiency and accuracy. 3. Investigate effective formats of AI-generated fact checks to enhance human trust through transparent explanations and evidence. |
| RQ3 | Limited Realistic, Complex, Multilingual Datasets | <ol style="list-style-type: none"> 1. Develop more realistic, complex, dynamic, domain-specific, multilingual fact-checking datasets with high-quality evidence for evaluation and fine-tuning. 2. Develop innovative and efficient data creation and annotation methodologies. 3. Develop more robust weak supervision, semi-supervised, or active learning techniques to reduce reliance on fully manual annotation. 4. Design systems and protocols for continuous data collection and dataset updates to reflect the real-time nature of information and misinformation. |
| RQ4 | Prompt Sensitivity and Adaptation Challenges | <ol style="list-style-type: none"> 1. Design and evaluate prompting methodologies that explicitly enforce and enable verification of evidence-grounded and faithful reasoning. 2. Develop adaptive, model-aware prompting frameworks that automatically generate and refine prompts and in-context examples to ensure robustness against variations. 3. Develop continual learning strategies for fine-tuned and domain-specific models to allow them to adapt to new information; investigate meta-learning or adaptive techniques for prompt optimization. 4. Develop prompting and fine-tuning methodologies that explicitly optimize for generating controllable, verifiable, and evidence-grounded reasoning and explanations. |
| RQ5 | Efficient Retrieval Explainable | <ol style="list-style-type: none"> 1. Design systems where LLMs can iteratively refine queries, explore multiple information angles, or retrieve evidence for decomposed sub-claims to build a more comprehensive evidence base. 2. Advanced agent-based RAG systems where an LLM (or multiple specialized LLM agents) can plan a sequence of reasoning and retrieval steps. 3. Design efficient RAG architectures that minimize computational overhead through optimized context chunking, selective retrieval, and reusable memory. |

and auditory information, and detecting sophisticated cross-modal manipulations is a key frontier [31, 43]. Equally vital are dedicated efforts to develop and evaluate information and effective and equitable fact-checking in a diverse spectrum of languages, with particular attention to resource-scarce linguistic contexts [32, 53].

Table 9 provides an overview of the identified gaps and future research agendas.

8 Conclusion

The inclusion of LLMs in automated fact-checking is changing the field. These models are reshaping how we process and verify the overwhelming amount of information we face online. Our work lays out the current research in this space, drawing attention to five critical areas: how LLMs are evaluated, the problem of hallucinations (where models produce false information), the importance of data sources, various ways of improving performance, and the growing use of RAG. Furthermore, in addition to extensive textual fact-checking, our review also acknowledges assessment of multimodal and multilingual dimensions, recognized as core areas of contemporary research. The findings highlight the fact that while this field is moving quickly and holds

considerable promise, there are significant challenges that still need careful attention if we want these systems to work reliably.

One of the major achievements of this study is that it pulls together a broad snapshot of what is going on right now. It highlights a tricky balance: on the one hand, LLMs have the potential to improve the speed and quality of fact-checking. However, these systems are still limited, and fixing those limits is going to require a lot of effort from researchers. It is essential to examine the practical utility of employing LLMs across all tasks, as smaller, task-specific models can often outperform them when properly optimized. From prior studies, one thing becomes clear: we need better tools to judge these systems. Since most existing tools only focus on whether a fact is right or wrong, there is now a growing need for tools that also look at how clearly the model explains things, whether it sticks to the logic, how well it shows where its answers came from, and how tough it is against tricky questions or misleading setups. From our review, it is evident that it is still difficult to ensure that LLMs always stick to real facts, especially when the situation is new or tricky. This improvement will require fundamental changes in the way these models are built and trained.

Although this review offers a comprehensive synthesis of

the existing literature, its scope is inherently constrained by the rapid velocity of technological innovation within the LLM domain. Consequently, emerging preprint findings, advancements in proprietary models, and developing best practices may not be fully encapsulated. Looking ahead, future research must prioritize the development of more sophisticated, robust, and universally standardized evaluation benchmarks. Such benchmarks are urgently needed to assess the factual accuracy, logical coherence, soundness of LLM reasoning, quality, utility, and persuasiveness of generated explanations, as well as the models' resilience to a wide array of adversarial attacks and evolving misinformation tactics.

The review presents key insights for developers, policy-makers, and all stakeholders who rely on online information. It underscores the need to address significant shortcomings in the development and evaluation of LLMs to enable their effective use as reliable fact-checking tools. That means paying close attention to how fair and balanced the data are, learning how to handle different languages and types of media better, and setting up strong rules and protections to make sure these systems are used ethically and do not cause harm.

Declarations

Conflict of Interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding: No external funding is available for this research.

Data Availability Statement: Not Applicable.

Ethics Approval and Consent to Participate. Not Applicable.

Informed Consents: Not Applicable.

Author Contributions: *Conceptualization and Methodology:* Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba, Mohaimenul Azam Khan Raiaan;

Resources and Literature Review: Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba;

Writing – Original Draft Preparation: Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba, Md. Abdur Rahman, Sadia Sultana Chowdhury, Mohaimenul Azam Khan Raiaan;

Validation: Sami Azam, Sadia Sultana Chowdhury, Mohaimenul Azam Khan Raiaan, Md. Abdur Rahman;

Formal Analysis: Md. Abdur Rahman, Sami Azam;

Writing – Reviewing and Finalization: Sami Azam, Mohaimenul Azam Khan Raiaan;

Project Supervision: Sami Azam, Mohaimenul Azam Khan Raiaan;

Project Administration: Sami Azam, Mohaimenul Azam Khan Raiaan.

References

- [1] D. Kampelopoulos, A. Tsanousa, S. Vrochidis, and I. Kompatsiaris, "A review of llms and their applications in the architecture, engineering and construction industry," *Artificial Intelligence Review*, vol. 58, no. 8, p. 250, 2025.
- [2] X. Wang, H. Jiang, Y. Yu, J. Yu, Y. Lin, P. Yi, Y. Wang, Y. Qiao, L. Li, and F.-Y. Wang, "Building intelligence identification system via large language model watermarking: a survey and beyond," *Artificial Intelligence Review*, vol. 58, no. 8, p. 249, 2025.
- [3] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy *et al.*, "Factuality challenges in the era of large language models and opportunities for fact-checking," *Nature Machine Intelligence*, vol. 6, no. 8, pp. 852–863, 2024.
- [4] T. Huang, "Content moderation by llm: From accuracy to legitimacy," *Artificial Intelligence Review*, vol. 58, no. 10, pp. 1–32, 2025.
- [5] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022, pp. 3214–3252.
- [6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.
- [7] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis) contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, 2021.
- [8] F. Ladhak, E. Durmus, M. Suzgun, T. Zhang, D. Jurafsky, K. McKeown, and T. B. Hashimoto, "When do pre-training biases propagate to downstream tasks? a case study in text summarization," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 3206–3219.
- [9] R. Aly, S. Papay, C. Christodoulopoulos, and I. Augenstein, "Feverous: Fact extraction and verification over unstructured and structured information," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 6118–6129.

- [10] L. Tang, P. Laban, and G. Durrett, “MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024, pp. 8818–8847.
- [11] V. Setty, “Surprising efficacy of fine-tuned transformers for fact-checking over larger language models,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2842–2846.
- [12] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi, “Bad actor, good advisor: Exploring the role of large language models in fake news detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 22105–22113.
- [13] L. Kupershstein, O. Zalepa, V. Sorokolit, and S. Prokopenko, “Ai-agent-based system for fact-checking support using large language models,” in *CEUR Workshop Proceedings*, 2025, pp. 321–331.
- [14] X. Zhao, L. Wang, Z. Wang, H. Cheng, R. Zhang, and K.-F. Wong, “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)*, 2024, pp. 12564–12573.
- [15] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen *et al.*, “Check your facts and try again: Improving large language models with external knowledge and automated feedback,” *arXiv preprint arXiv:2302.12813*, 2023.
- [16] J. Ma, L. Hu, R. Li, and W. Fu, “Local: Logical and causal fact-checking with llm-based multi-agents,” in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 1614–1625.
- [17] B. Ghosh, S. Hasan, N. A. Arifat, and A. Khan, “Logical Consistency of Large Language Models in Fact-Checking,” *The Thirteenth International Conference on Learning Representations*, 2025.
- [18] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, “A unified llm-kg framework to assist fact-checking in public deliberation,” in *Proceedings of the First Workshop on Language-Driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, 2024, pp. 13–19.
- [19] J. Kasai, K. Sakaguchi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, K. Inui *et al.*, “Realtme qa: What’s the answer right now?” *Advances in neural information processing systems*, vol. 36, pp. 49025–49043, 2023.
- [20] M. Li, B. Peng, M. Galley, J. Gao, and Z. Zhang, “Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 163–181. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-naacl.12>
- [21] X. Zhang and W. Gao, “Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Nusa Dua, Bali, 2023, pp. 996–1011.
- [22] M. A. Khaliq, P. Y. Chang, M. Ma, B. Pflugfelder, and F. Miletic, “RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models,” in *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Miami, Florida, USA, 2024, pp. 280–296.
- [23] I. Vykopal, M. Pikuliak, S. Ostermann, and M. Šimko, “Generative large language models in automated fact-checking: A survey,” *arXiv preprint arXiv:2407.02351v2*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.02351>
- [24] A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein, “Claim verification in the age of large language models: A survey,” *arXiv preprint arXiv:2408.14317v2*, 2025. [Online]. Available: <https://arxiv.org/abs/2408.14317>
- [25] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, and P. Nakov, “Factuality of Large Language Models: A Survey,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024, pp. 19519–19529.
- [26] B. A. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. G. Linkman, “Systematic literature reviews in software engineering - A systematic literature review,” *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009. [Online]. Available: <https://doi.org/10.1016/j.infsof.2008.09.009>
- [27] M. E. Conway, “How do committees invent,” *Datamation*, vol. 14, no. 4, pp. 28–31, 1968.
- [28] B. Zhang, A. Bornet, A. Yazdani, P. Khlebnikov, M. Milutinovic, H. Rouhizadeh, P. Amini, and D. Teodoro, “A dataset for evaluating clinical research claims in large language models,” *Scientific Data*, vol. 12, no. 1, p. 86, 2025.

- [29] R. Kamoi, S. S. S. Das, R. Lou, J. J. Ahn, Y. Zhao, X. Lu, N. Zhang, Y. Zhang, H. R. Zhang, S. R. Vummanthala, S. Dave, S. Qin, A. Cohan, W. Yin, and R. Zhang, “Evaluating LLMs at Detecting Errors in LLM Responses,” in *First Conference on Language Modeling*, 2024.
- [30] H. Tran, J. Wang, Y. Ting, W. Huang, and T. Chen, “Leaf: Learning and evaluation augmented by fact-checking to improve factualness in large language models,” *arXiv preprint arXiv:2410.23526*, 2024.
- [31] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, “Sniffer: Multimodal large language model for explainable out-of-context misinformation detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 13 052–13 062.
- [32] A. Singhal, T. Law, C. Kassner, A. Gupta, E. Duan, A. Damle, and R. L. Li, “Multilingual fact-checking using llms,” in *Proceedings of the Third Workshop on NLP for Positive Impact*, 2024, pp. 13–31.
- [33] C. Si, N. Goyal, T. Wu, C. Zhao, S. Feng, H. Daume III, and J. Boyd-Graber, “Large language models help humans verify truthfulness – except when they are convincingly wrong,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, 2024, pp. 1459–1474. [Online]. Available: <https://aclanthology.org/2024.naacl-long.81/>
- [34] Z. Xie, R. Xing, Y. Wang, J. Geng, H. Iqbal, D. Sahnan, I. Gurevych, and P. Nakov, “FIRE: Fact-checking with Iterative Retrieval and Verification,” in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, 2025, pp. 2901–2914.
- [35] N. Lee, B. Z. Li, S. Wang, W. Yih, H. Ma, and M. Khabsa, “Language Models as Fact Checkers?” in *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Online, 2020, pp. 36–41.
- [36] H. Cao, L. Wei, W. Zhou, and S. Hu, “Multi-source knowledge enhanced graph attention networks for multimodal fact verification,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [37] Y. Liu, H. Sun, W. Guo, X. Xiao, C. Mao, Z. Yu, and R. Yan, “Bidev: Bilateral defusing verification for complex claim fact-checking,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, 2025, pp. 541–549.
- [38] J. Vladika, I. Hacajova, and F. Matthes, “Step-by-Step Fact Verification System for Medical Claims with Explainable Reasoning,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, Albuquerque, New Mexico, 2025, pp. 805–816.
- [39] H. Sankararaman, M. N. Yasin, T. Sorensen, A. Di Bari, and A. Stolcke, “Provenance: A Lightweight Fact-checker for Retrieval Augmented LLM Generation Output,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Miami, Florida, US, 2024, pp. 1305–1313.
- [40] J. Magomere, E. La Malfa, M. Tonneau, A. Kazemi, and S. A. Hale, “When Claims Evolve: Evaluating and Enhancing the Robustness of Embedding Models Against Misinformation Edits,” *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22 374–22 404, 2025.
- [41] D. Pisarevskaya and A. Zubia, “Zero-shot and Few-shot Learning with Instruction-following LLMs for Claim Matching in Automated Fact-checking,” in *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, 2025, pp. 9721–9736.
- [42] V. Krishnamurthy and V. Balaji, “Yours truly: A credibility framework for effortless llm-powered fact checking,” *IEEE Access*, 2024.
- [43] Y. Ge, X. Zeng, J. S. Huffman, T.-Y. Lin, M.-Y. Liu, and Y. Cui, “Visual fact checker: enabling high-fidelity detailed caption generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 033–14 042.
- [44] Y. Wang, M. Wang, H. Iqbal, G. N. Georgiev, J. Geng, I. Gurevych, and P. Nakov, “Openfactcheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and llms,” in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 11 399–11 421.
- [45] H. Lin, Y. Deng, Y. Gu, W. Zhang, J. Ma, S. Ng, and T. Chua, “FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025, pp. 360–381.
- [46] T.-H. Cheung and K.-M. Lam, “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 (APSIPA ASC)*. IEEE, 2023, pp. 846–853.
- [47] E. C. Choi and E. Ferrara, “Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation,” in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1441–1449.
- [48] D. Quelle and A. Bovet, “The perils and promises of fact-checking with large language models,” *Frontiers in Artificial Intelligence*, vol. 7, p. 1341697, 2024.
- [49] X. Zhao, J. Yu, Z. Liu, J. Wang, D. Li, Y. Chen, B. Hu, and M. Zhang, “Medico: Towards Hallucination Detection and Correction with Multi-source Evidence Fusion,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Miami, Florida, USA, 2024, pp. 34–45.
- [50] X. Jing, S. Billa, and D. Godbout, “On A Scale From 1 to 5: Quantifying Hallucination in Faithfulness Evaluation,” in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, 2025, pp. 7765–7780.
- [51] X. Zhang and W. Gao, “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 (LREC-COLING 2024)*, Torino, Italia, 2024, pp. 13 861–13 873.
- [52] M. Schlichtkrull, Z. Guo, and A. Vlachos, “Averitec: A dataset for real-world claim verification with evidence from the web,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 128–65 167, 2023.
- [53] R. Singal, P. Patwa, P. Patwa, A. Chadha, and A. Das, “Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs,” in *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Miami, Florida, USA, 2024, pp. 91–98.
- [54] V. Chatrath, M. Lotif, and S. Raza, “Fact or Fiction? Can LLMs be Reliable Annotators for Political Truths?” in *Workshop on Socially Responsible Language Modelling Research*, 2024.
- [55] J. Leite, O. Razuvayevskaya, K. Bontcheva, and C. Scarton, “Weakly Supervised Veracity Classification with LLM-Predicted Credibility Signals,” 2024, <https://arxiv.org/abs/2309.07601>.
- [56] Y. Ding, M. Facciani, E. Joyce, A. Poudel, S. Bhattacharya, B. Veeramani, S. Aguinaga, and T. Weninger, “Citations and trust in llm generated responses,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 787–23 795.
- [57] J. Geng, Y. Kementchedjhieva, P. Nakov, and I. Gurevych, “Multimodal large language models to support real-world fact-checking,” *arXiv preprint arXiv:2403.03627*, 2024.
- [58] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, “A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity,” *arXiv preprint arXiv:2302.04023*, 2023.
- [59] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Hadidow, A. Birch, P. Colombo, and A. F. Martins, “Hallucinations in large multilingual translation models,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023.
- [60] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.
- [61] T. Sakurai, S. Shiramatsu, and R. Kinoshita, “Llm-based agent for recommending information related to web discussions at appropriate timing,” in *2024 IEEE International Conference on Agents (ICA)*. IEEE, 2024, pp. 120–123.
- [62] Y. Huang, X. Feng, X. Feng, and B. Qin, “The factual inconsistency problem in abstractive text summarization: A survey,” *arXiv preprint arXiv:2104.14839*, 2021.
- [63] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [64] W. Li, W. Wu, M. Chen, J. Liu, X. Xiao, and H. Wu, “Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods,” *arXiv preprint arXiv:2203.05227*, 2022.
- [65] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu *et al.*, “Lima: Less is more for alignment,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 006–55 021, 2023.

- [66] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, “Aligning large language models with human: A survey,” *arXiv preprint arXiv:2307.12966*, 2023.
- [67] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, “Ethical and social risks of harm from language models,” *arXiv preprint arXiv:2112.04359*, 2021.
- [68] D. Li, A. S. Rawat, M. Zaheer, X. Wang, M. Lukasik, A. Veit, F. Yu, and S. Kumar, “Large language models with controllable working memory,” *arXiv preprint arXiv:2211.05110*, 2022.
- [69] Y. Onoe, M. J. Zhang, E. Choi, and G. Durrett, “Entity cloze by date: What lms know about unseen entities,” *arXiv preprint arXiv:2205.02832*, 2022.
- [70] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, Y.-Y. Liu, and L. Yuan, “Llm lies: Hallucinations are not bugs, but features as adversarial examples,” *arXiv preprint arXiv:2310.01469*, 2023.
- [71] X. Li, Y. Zhang, and E. C. Malthouse, “Large language model agent for fake news detection,” *arXiv preprint arXiv:2405.01593*, 2024.
- [72] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer, “Fact-checking information from large language models can decrease headline discernment,” *Proceedings of the National Academy of Sciences*, vol. 121, no. 50, p. e2322823121, 2024.
- [73] G. Luo, T. Darrell, and A. Rohrbach, “NewsCLIPings: Automatic Generation of Out-of-Context Multimodal Media,” in *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*, 2021, pp. 6801–6817. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.545/>
- [74] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, and Q. V. Le, “Long-form factuality in large language models,” *arXiv*, 2024.
- [75] M. Leippold, S. A. Vaghefi, D. Stammbach, V. Muccione, J. Bingler, J. Ni, C. C. Senni, T. Wekhof, T. Schimanski, G. Gostlow *et al.*, “Automated fact-checking of climate claims with large language models,” *npj Climate Action*, vol. 4, no. 1, p. 17, 2025.
- [76] E. Fadeeva, A. Rubashevskii, A. Shelmanov, S. Petrakov, H. Li, H. Mubarak, E. Tsymbalov, G. Kuzmin, A. Panchenko, T. Baldwin *et al.*, “Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification,” in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024, pp. 9367–9385.
- [77] B. M. Yao, A. Shah, L. Sun, J. Cho, and L. Huang, “End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, 2023, pp. 2733–2743.
- [78] S. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “RED-DOT: Multimodal Fact-Checking via Relevant Evidence Detection,” *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2025.
- [79] P. Sharma, T. R. Shaham, M. Baradad, S. Fu, A. Rodriguez-Munoz, S. Duggal, P. Isola, and A. Torralba, “A vision check-up for language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14410–14419.
- [80] J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu, “LLM-Enhanced multimodal detection of fake news,” *PLoS one*, vol. 19, no. 10, p. e0312240, 2024.
- [81] K. Kakizaki, Y. Matsunaga, and R. Furukawa, “MAFT: Multimodal Automated Fact-Checking via Textualization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, pp. 29646–29648.
- [82] X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak, “Don’t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 7915–7927.
- [83] S. Shafayat, E. Kim, J. Oh, and A. Oh, “Multi-FAct: Assessing Factuality of Multilingual LLMs using FActScore,” in *First Conference on Language Modeling (COLM)*, 2024.
- [84] M. Siino, E. Di Nuovo, I. Tinnirello, and M. La Cascia, “Fake News Spreaders Detection: Sometimes Attention Is Not All You Need,” *Inf.*, vol. 13, no. 9, p. 426, 2022.
- [85] H. Shcharbakova, T. Anikina, N. Skachkova, and J. V. Genabith, “When Scale Meets Diversity: Evaluating Language Models on Fine-Grained Multilingual Claim Verification,” in *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, Vienna, Austria, 2025, pp. 69–84.

- [86] S. Z. Jannah, E. Aco, S. Peng, S. Wakamiya, and E. Aramaki, “Multilingual Symptom Detection on Social Media: Enhancing Health-related Fact-checking with LLMs,” in *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, Vienna, Austria, 2025, pp. 54–68.
- [87] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, “Benchmarking Retrieval-Augmented Generation for Medicine,” in *ACL Anthology*, 2024, pp. 6233–6251.
- [88] E. C. Choi and E. Ferrara, “Fact-gpt: Fact-checking augmentation via claim matching with llms,” in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 883–886.
- [89] R. T. Sairaj and S. Balasundaram, “Ontology Mapping for Retrieval Augmented Modelling to Reduce Factual Hallucinations in Pretrained Language Model-Based Auto-Generated Questions,” *Applied Ontology*, p. 15705838251343009, 2025.
- [90] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, “Multi-modal fusion with recurrent neural networks for rumor detection on microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.
- [91] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, “Mdfend: Multi-domain fake news detection,” in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 3343–3347.
- [92] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media,” *arXiv*, 2018.
- [93] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, “Fa-kes: A fake news dataset around the syrian war,” in *Proceedings of the international AAAI conference on web and social media*, vol. 13, 2019, pp. 573–582.
- [94] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, and C. Scarton, “EUvsDisinfo: A dataset for multilingual detection of pro-Kremlin disinformation in news articles,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 5380–5384.
- [95] Q. Yang, T. Christensen, S. Gilda, J. Fernandes, D. Oliveira, R. Wilson, and D. Woodard, “Are fact-checking tools helpful? an exploration of the usability of google fact check,” in *International Conference on Data and Information in Online*, 2024, pp. 82–98.
- [96] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “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*, 2018, pp. 809–819. [Online]. Available: <https://aclanthology.org/N18-1074/>
- [97] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, “Hover: A dataset for many-hop fact extraction and claim verification,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3441–3450. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.309/>
- [98] W. Y. Wang, ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 422–426. [Online]. Available: <https://aclanthology.org/P17-2067/>
- [99] Z. Yang *et al.*, “A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection,” in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 2637–2647. [Online]. Available: <https://aclanthology.org/2022.coling-1.230/>
- [100] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, “Fact or Fiction: Verifying Scientific Claims,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7534–7550. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.609/>
- [101] A. Saakyan, T. Chakrabarty, and S. Mureşan, “COVID-FACT: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 2116–2129. [Online]. Available: <https://aclanthology.org/2021.acl-long.165/>
- [102] A. Pal, L. K. Umaphathi, and M. Sankarasubbu, “MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering,” in *Proceedings of the Conference on Health, Inference, and Learning*, 2022, pp. 248–260. [Online]. Available: <https://proceedings.mlr.press/v174/pal22a.html>
- [103] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos *et al.*,

- “An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015.
- [104] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, “PubMedQA: A Dataset for Biomedical Research Question Answering,” 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)*, 2019, pp. 2567–2577. [Online]. Available: <https://aclanthology.org/D19-1259/>
- [105] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. N. Reganti, P. Patwa, A. Das, T. Chakraborty, A. P. Sheth, A. Ekbal *et al.*, “FACTIFY: A Multi-Modal Fact Verification Dataset.” in *DE-FACTIFY@ AAAI*, 2022.
- [106] B. Yao, A. Shah, L. Sun, J. Cho, and L. Huang, “End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models,” in *Proceedings Of The 46th International ACM SIGIR Conference On Research And Development In Information Retrieval*, 2023, pp. 2733–2743. [Online]. Available: <https://doi.org/10.1145/3539618.3591879>
- [107] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4567–4578. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.397/>
- [108] R. Kamoi, S. S. S. Das, R. Lou, J. J. Ahn, Y. Zhao *et al.*, “Evaluating LLMs at Detecting Errors in LLM Responses,” *arXiv preprint arXiv:2404.03602*, 2024. [Online]. Available: <https://arxiv.org/abs/2404.03602>
- [109] J. M. Eisenschlos, B. Dhingra, J. Bulian, B. Börschinger, and J. Boyd-Graber, “Fool Me Twice: Entailment from Wikipedia Gamification,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021, pp. 1234–1245. [Online]. Available: <https://aclanthology.org/2021.naacl-main.32/>
- [110] F. Bayat, L. Zhang, S. Munir, and L. Wang, “FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation,” 2025, <https://arxiv.org/abs/2410.22257>.
- [111] P. Li, Z. Gao, B. Zhang, T. Yuan, Y. Wu, M. Harandi, Y. Jia, S. Zhu, and Q. Li, “FIRE: A Dataset for Feedback Integration and Refinement Evaluation of Multimodal Models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.11522>
- [112] R. Panchendarajan, R. Míguez, and A. Zubíaga, “MultiClaimNet: A massively multilingual dataset of fact-checked claim clusters,” *arXiv preprint arXiv:2503.22280*, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2503.22280>
- [113] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struš, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani *et al.*, “Overview of the CLEF–2022 CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2022, pp. 495–520.
- [114] A. Gupta and V. Srikumar, “X-Fact: A New Benchmark Dataset for Multilingual Fact Checking,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 732–748. [Online]. Available: <https://doi.org/10.48550/arXiv.2106.09248>
- [115] M. Siino, “BrainLlama at SemEval-2024 Task 6: Prompting Llama to detect hallucinations and related observable overgeneration mistakes,” in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico, 2024, pp. 82–87.
- [116] M. Siino and I. Tinnirello, “GPT hallucination detection through prompt engineering,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.
- [117] R. T. Sairaj and S. R. Balasundaram, “Ensemble learning with RAG model to reduce redundant question topics in auto-generated exam questions,” *Discover Computing*, vol. 28, no. 1, pp. 1–17, 2025.
- [118] Y. Bai and K. Fu, “A large language model-based fake news detection framework with rag fact-checking,” in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8617–8619.
- [119] S. S. Chowdhury, R. Alvi, S. S. Rahman, M. A. Rahman, M. A. K. Raiaan, M. R. Islam, M. Hussain, and S. Azam, “From language to action: A review of large language models as autonomous agents and tool users,” *arXiv preprint arXiv:2508.17281*, 2025.
- [120] A. I. Abian, M. A. K. Raiaan, A. Karim, S. Azam, N. M. Fahad, N. Shafibady, K. C. Yeo, and F. De Boer, “Automated diagnosis of respiratory diseases from lung ultrasound videos ensuring XAI: an innovative hybrid model approach,” *Frontiers in Computer Science*, vol. 6, 2024.

- [121] A. I. Abian, M. A. K. Raiaan, M. Jonkman, S. M. S. Islam, and S. Azam, “Atrous spatial pyramid pooling with swin transformer model for classification of gastrointestinal tract diseases from videos with enhanced explainability,” *Engineering Applications of Artificial Intelligence*, vol. 150, p. 110656, 2025.
- [122] L. Sheng, F. Chang, Q. Sun, D. Wangzha, and Z. Gu, “Uncertainty reports as explainable ai: A cognitive-adaptive framework for human-ai decision systems in context tasks,” *Advanced Engineering Informatics*, vol. 68, p. 103634, 2025.
- [123] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, P. E. Ukeoma, and N. F. Dibiaezue, “Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of ai models for human,” *Neural Processing Letters*, vol. 57, no. 1, p. 16, 2025.
- [124] J. C. Cheung and S. S. Ho, “The effectiveness of explainable ai on human factors in trust models,” *Scientific Reports*, vol. 15, no. 1, p. 23337, 2025.