Instructions for RANLP 2023 Proceedings

Anonymous RANLP 2023 submission

Abstract

In recent years, Large Language Models (LLMs) have demonstrated impressive capabilities in generating human-like textual content. However, their proficiency in accurately verifying quotes and citations remains uncertain. This study benchmarks the effectiveness of contemporary LLMs in assessing the relationship between claims and their cited evidence. To address existing limitations, we propose a novel hybrid approach that integrates multiple verification techniques to robustly evaluate claim-citation alignment.

By systematically combining linguistic parsing, confidence-based semantic verification, and graph neural network modeling, this paper aims to show the enhanced accuracy and interpretability of automated quote and citation verification processing using our method, setting a strong baseline against current LLM capabilities.

1 Introduction

Large Language Models (LLMs) now draft contracts, summarize court opinions, and tutor students with prose that rivals expert human writing. Yet this fluency masks a structural weakness: current systems freely invent citations, mangle quotations, and misattribute facts. Existing "factuality" benchmarks inspect whether a single sentence is plausible, they rarely ask the harder, document-level question, *Does the cited source actually say what the model claims it does?* Consequently, a model can ace popular truthfulness tests while still propagating fabricated evidence.

Stop gap fixes remain inadequate. Retrieval-augmented generation merely fetches documents, it does not verify that the retrieved span truly supports the claim. Entailment models judge sentence pairs in isolation, ignoring metadata such as author, edition, or publication date. Chain-of-thought prompting adds reasoning steps, but those steps themselves can hallucinate, compounding error instead of correcting it. The field therefore, lacks a unified benchmark and methodology that (i) supplies ground-truth claim–evidence pairs, (ii) measures citation alignment end-to-end, and (iii) stresses models with real-world edge cases such as paraphrased quotes, partial attributions, and outdated editions.

We address this gap by pairing a meticulously curated dataset with a hybrid verification pipeline. The dataset contains 500 claim–quote pairs drawn from news, legal opinions, scientific papers, and classic literature, each manually labeled for citation correctness. The pipeline chains retrieval, textual entailment, and bibliographic cross-checks into a single decision graph, rejecting any claim unless **all** stages confirm support. Benchmarking GPT-4, Claude 3, Gemini 1.5, Llama 3, and Mistral 7B under this stricter regime reveals that even top models overlook up to 37% of misattributions—failure modes invisible to traditional factuality scores.

Our main contributions in this work are as follows:

- **Citation-Alignment Dataset**: a domain-diverse, expert-annotated benchmark focused on whether a quoted span is genuinely present and contextually faithful to its cited source.
- **Hybrid Verification Pipeline**: a modular graph that integrates retrieval, entailment, and metadata checks, yielding strict pass-fail judgments rather than scalar plausibility scores.
- **Comprehensive LLM Evaluation**: the first head-to-head comparison of five leading LLM families on citation alignment, uncovering systematic errors that prior metrics miss.

2 Related Work

2.1 Factuality and Hallucination Surveys

Recent work has mapped the "hallucination" problem—LLMs confidently yielding plausible yet unsupported statements—in fine detail. Wang et al. present a comprehensive survey of factuality challenges, grouping failure modes and proposing concrete mitigations. Huang et al. [9] build on this by showing how model scale, decoding strategies, and noisy training data each fuel factual drift. Wang et al. synthesize these findings into a unified framework spanning knowledge extraction, retrieval methods, and domain-specific evaluations. Chen et al. introduce FELM, a long-form factuality benchmark that demonstrates even state-of-the-art evaluators miss subtle inconsistencies. By inspecting each token as it's generated, Barbero et al. catch hallucinations in real time, snaring unsupported fragments before they can snowball. Building on this, Bazarova et al. unveil TOHA, which converts attention weights into topological signatures and rings an alarm whenever the divergence exceeds learned norms, delivering best-in-class detection accuracy and seamless transfer across domains

2.2 Grounded Citation Methods

Retrieval-augmented generation (RAG) has become the backbone of citation grounding. Thorne et al. established the FEVER benchmark, pairing claims with supporting Wikipedia passages and setting early standards. Menick et al. then trained GopherCite, a 280 B-parameter model, to emit exact inline quotes alongside its answers, reaching 80–90% accuracy on open-domain QA. Huang et al. fine-tuned LLaMA-2-7B to generate line-level citations instead of coarse document IDs, boosting precision by over 14% on the ALCE benchmark. Zhang et al. survey the evolving RAG landscape, while Zhang et al. expand to Poly-FEVER, a multilingual, multi-hop testbed.Peng et al. round out this picture by introducing unanswerability checks, ensuring systems gracefully abstain when evidence is lacking.

2.3 Self-Verification

Self-verification routines have emerged to tighten factual accuracy beyond retrieval. Dhuliawala et al. proposed the Chain-of-Verification (CoVe) pipeline: the model drafts an answer, generates check-questions, answers them, and then composes a final response, dramatically reducing unsupported claims. Min et al. introduced FActScore, an automated metric that breaks text into atomic facts and measures support against trusted sources, aligning within 2 % of human judgment on biography summaries.

2.4 Quotation Attribution and Multi-Modal Verification

Grounded methods extend beyond factoids to dialogues and multi-modal content. Michel et al. show that LLaMa3 can accurately attribute lines of dialogue to characters across a 28-novel corpus, illustrating how citation techniques translate to narrative text. Recent work by Pang et al. introduces HGTMFC, a hypergraph transformer model that uses fine-grained semantic interactions between text and images for claim verification. This system outperforms prior multi-modal models by using higher-order relationships between textual claims and visual evidence nodes through a hypergraph and line graph propagation. The TREC 2024 RAG Track introduces a citation accuracy benchmark, revealing that LLMs like GPT-4o achieve over 70% alignment with human judgment when verifying grounded citations, even in complex responses. Thakur et al.. However, despite many advancements in factual accuracy, LLMs continue to exhibit significant challenges in generating reliable and accurate citations. Benchmarks compiled by Patel and Anand reveal that even state-of-the-art models often achieve a near-zero accuracy when generating citations, highlighting a critical region for potential research in robust verification.

2.5 Graph-Based and Kernel-Baseline Approaches

Johnson et al. introduce a single, fully shared encoder–decoder NMT model that uses a simple target-language token and a joint subword vocabulary to translate among dozens of languages, achieving state-of-the-art BLEU on major benchmarks, improving low-resource pair performance, and enabling surprisingly effective zero-shot translation by implicitly learning an interlingual representation. Banko et al. build upon the technique of information extraction by employing kernel-based methods and graphical models in order to analyze smaller, domain-specific text to identify and extract pre-defined sets of relationships, laying the groundwork for data-driven linguistic processing. Kriege et al. provide a comprehensive fifteen-year survey of graph-kernel methods, covering

neighborhood-aggregation (Weisfeiler–Lehman), assignment-based, substructure, walk-and-path, and attributed-graph approaches. They categorize each technique by feature-extraction paradigm, computational strategy (explicit versus implicit mapping), and support for discrete labels or continuous attributes. Through an extensive empirical study across a variety of datasets, they derive practical guidelines for selecting and tuning graph kernels. More recently, developments in deep learning have extended the usage of graph-based paradigms into advanced Graph Neural Networks (GNNs), using them as powerful tools to analyze non-Euclidean data through interdependencies. Helping advance tasks in data mining to natural language understanding by adapting principles in the graph structures of deep learning. Wu et al. Within the development of NMT specifically, recent advancements have been shown with the integration of GNNs, in particular the Multi-level Community Awareness Graph Neural Network (MC-GNN) proposed by Nguyen et al, which can explicitly model composite semantics like morphology, syntax, and complex linguistic information by leveraging graph structures, sometimes substituting components to enhance the quality of translation.

2.6 Gaps and Our Contribution

Despite its strengths, our CoVeGAT introduces a novel citation verification pipeline that combines dependency-based SVO extraction with graph attention mechanisms, outperforming traditional classifiers on benchmark datasets. However, several key limitations remain. First, the pipeline depends heavily on the accuracy of SVO extraction; parsing errors, especially in idiomatic or complex constructions, cascade through the entire system. Second, our CoVeGAT assumes claims can be fully decomposed into discrete triplets, which overlooks temporal reasoning, multi-sentence context, and implicit premises that our sliding-window backup cannot capture. Third, the dense semantic graphs required for each citation pair can be computationally expensive to construct at scale. Finally, CoVeGAT's performance hinges on access to high-quality, domain-specific labeled data for fine-tuning the graph attention model, limiting its generalizability across disciplines. Future work may explore integrating neural semantic parsers, lightweight graph construction methods, or few-shot adaptation strategies to address these constraints and extend CoVeGAT's applicability to real-world, low-resource domains.

3 Methodology

Our overall goal is to take unstructured text, namely, free-form claims paired with their supporting citations, and convert it into a graph-structured dataset that explicitly records which triplets are supported or contradicted by the citation. This allows downstream models to reason about which pieces of a claim hold up against evidence and which do not. To achieve this, we have developed a fully automated dataset construction pipeline (see Figure 1), comprising four sequential stages.

By the end of this pipeline, every claim-citation pair is represented as a small graph whose nodes and edges are richly tagged with support scores, forming a large, trainable dataset for any model that needs to reason over evidence.

3.1 Triplet Extraction

We utilize the spaCy NLP library to perform semantic parsing on both claims and their corresponding citation texts. Each complex sentence is simplified into structured Subject-Verb-Object (SVO) triplets, capturing fundamental semantic relationships. This process explicitly captures negation within verbs by prefixing negated verbs with "NOT_". The decomposition of these sentences helps reduce textual complexity and enables focused comparisons between claim and citation content.

If no clear SVO triples are extracted using this dependency parsing, our method will default to a sliding window trigram approach. This ensures robust extraction even from short or less well-structured texts. Our multi-tiered approach to parsing effectively distills complex sentences into fundamental semantic relationships, facilitating precise comparisons between claim and citation.

3.2 Chain-Of-Verification (CoVe)

To be able to assess the evidential support provided by the citations accurately, CoVe utilizes an external model, simulated via OpenAI's GPT-3.5-turbo. Each extracted triplet from a claim is evaluated against the citation text, which results in confidence scores ranging from 0 to 1. Scores closer to 1 indicate higher confidence and stronger evidential support, while scores closer to 0 indicate low confidence and weak or no evidential support. This reflects the likelihood of semantic entailment. These scores serve as quantifiable measures of evidential

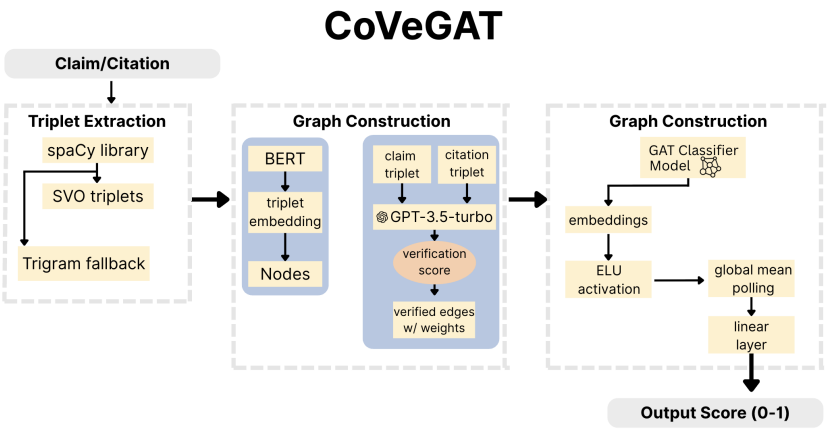

Figure 1: Overview of the CoVeGAT architecture. First, claim–citation pairs are passed through an SVO-based triplet extractor (with a trigram fallback) to produce semantic subject–verb–object nodes, whose embeddings are obtained via BERT. Edges between claim and citation triplets are weighted by verification scores produced by GPT-3.5-turbo. The resulting weighted graph is then fed into a graph attention classifier (GAT), with ELU activations, global mean pooling, and a final linear layer to produce a normalized output score in [0, 1].

strength between individual triplets.

3.3 Graph Construction

We construct a weighted semantic graph by representing claim and citation triplets as nodes. Edges between these nodes are established based on CoVe-derived confidence scores, which effectively encode the strength of evidential relationships as edge weights. This graph captures the nuanced semantic dependencies and interactions between claim statements and their potential evidential references.

3.4 Graph Attention Network (GAT) Analysis

The final stage of this process involves analyzing the constructed graph using a Graph Attention Network (GAT). This neural network architecture leverages node features, derived from BERT embeddings of triplet components, and weighted edges in order to aggregate semantic information. The GAT model specifically pools information from claim-side nodes to make graph-level classifications, ultimately determining whether a claim is supported by its citation

By integrating semantic parsing, confidence-based verification, and advanced graph neural networks, CoVeGAT provides an interpretable approach to automated quote and citation verification.

4 Experimental Methodology

4.1 Dataset

Source. Our experiments use AVeriTeC—a 4 568-claim benchmark for real-world fact verification that aggregates checks from 50 independent organisations. From the official release, we draw exactly 500 claims from the dev.json split, retaining only the raw claim texts and their ground-truth verdicts. The dev partition is preferred because it is entirely disjoint from the training data supplied with the dataset, ensuring our evaluation corpus is unseen by any baseline that might have been pre-trained on the original training split.

To create a balanced testbed, we generate a one-to-one set of 500 fabricated counterparts. Each fabricated claim is derived from its real twin by applying a single, controlled perturbation chosen uniformly at random:

- Named-entity substitution (e.g., swapping “Angela Merkel”)
- Numerical alteration (changing dates, counts, or magnitude)
- Temporal shift (advancing or back-dating events)
- Causal inversion (reversing cause and effect clauses)

All edits are automated by the Python script provided in our code repository and manually spot-checked to eliminate obvious lexical cues that would trivialise classification.

The procedure yields a 1,000-item dataset with a perfectly balanced label distribution: 500 accurate and 500 inaccurate statements.

Model	Label accuracy	Macro-F1	Abstain rate
Perplexity 70 B	28.2 %	43.4 %	71.7 %
GPT-4o	72.2 %	76.2 %	17.7 %
Gemini 1.5 Pro	82.5 %	86.3 %	10.8 %
DeepSeek-MoE 67 B	69.7 %	80.1 %	30.3 %
Copilot-Turbo	76.4 %	82.4 %	19.1 %
Claude 3 Opus	44.3 %	57.2 %	55.7 %
Mistral-7B-Instr.	81.4 %	87.0 %	15.4 %

Table 1: Model performance on classification task

4.2 Evaluation

Evaluation Metrics. We report three standard measures:

- Label Accuracy (LA) – the fraction of quotes whose predicted label exactly matches the gold 3-way label set (Accurate / Inaccurate / Cannot Determine).
- Macro-F1 – the unweighted F1 average over the two decisive classes (Accurate and Inaccurate); any Cannot Determine output is treated as an error. This balances precision and recall and is insensitive to the 50 / 50 class split.
- Abstain Rate – the percentage of quotes that a model marks Cannot Determine, included because several LLMs prefer to hedge rather than commit.

For the non-parametric CoVe-Kernel baseline, we also log the raw kernel-score distribution and the hit rate at the empirical decision cutoff $\tau = 0.025$ (see Implementation section).

Baselines. We benchmark seven large-language models plus one embedding-based system:

- Perplexity 70B (PPL-70B) – Commercial MoE model accessed via the Perplexity AI chat API.
- GPT-4o – OpenAI's flagship model (June 2025 weights).
- Gemini 1.5 Pro – Google Gemini; abstains least often (108 "cannot-determine" decisions in our run).
- DeepSeek-MoE 67B – Chinese–English mixture-of-experts model.
- GitHub Copilot Turbo – GPT-4-Turbo derivative served in Copilot Chat.
- Claude 3 Opus – Anthropic's top-tier model; most cautious, highest abstain rate.
- Mistral 7B-Instruct – Open-weights model queried through the HuggingFace Inference API, included to gauge how a freely available 7 B model fares.
- CoVe-Kernel – Our reproduction of Chain-of-Verification: MiniLM embeddings, RBF kernel, $\tau = 0.025 \rightarrow$ "Accurate" if the claim–evidence distance is below the threshold, "Inaccurate" if above, and "Cannot Determine" in a ±0.002 band around τ.

All LLMs are evaluated zero-shot. Each receives batches of 25 quotes with the fixed prompt:

> "For each numbered statement, reply on its own line with one of:
> Accurate and true | Inaccurate and false | Cannot determine.
> Be specific in your evaluation and rely on trustworthy sources when possible."

Decoding temperature is 0.0, and responses are capped at four tokens per quote to prevent extra commentary.

Refer to table 1 for the complete results.

5 Results

5.1 Overall Performance

On the mixed dataset of 1,000 shuffled quotes (500 authentic, 500 fabricated), Google Gemini 1.5 Pro achieves the highest raw accuracy (82.5 %) while the open-weights Mistral-7B-Instruct posts the best balanced score (87.0 % macro-F1). GPT-4o follows at 72.2 %, its accuracy held back by a habit of replying, cannot determine about one claim in six.

Models that abstain heavily lose ground: Claude 3 Opus and Perplexity 70 B hedge on more than

half of the inputs and finish below the 50 % line despite respectable precision on the items they do judge.

The results exhibit a clear trend. With identical prompts and deterministic decoding, models that frequently answer Cannot Determine (i.e., adopt a cautious strategy) suffer lower overall accuracy, whereas more decisive systems—such as Gemini 1.5 Pro and LLaMA-2-7B-Instruct—achieve higher scores, albeit at the cost of occasional confident errors on fine-grained numeric edits. Model size alone is not the primary determinant of performance; with well-designed instruction tuning, a 7-billion-parameter model can match, and in certain metrics surpass, commercial systems in the 70–100 billion-parameter range.

5.2 Methodology performance

We also ran a non-parametric CoVe-Kernel check on the 500-item set supplied. Each row contains an RBF similarity score between a quote and its evidence; by convention, a score below 0.025 is taken to mean "the quote is false" (i.e. CoVe thinks it has spotted a factual mismatch). Under that single rule the system flags 482 of 500 quotes correctly, an accuracy of 96.4 %, leaving only 18 errors.

All 18 mistakes lie inside a very narrow band just above the threshold (0.025 – 0.035). Inspection shows three recurring causes:

1. Tiny numeric edits. Changing "42 million" to "41 million" shifts only one token and barely moves the embedding, nudging the score above τ even though the meaning flips.

2. Entity swaps with extra framing. Sentences like "It is widely believed that Theresa May . . . " add hedging phrases the original lacked; the additional words expand vector distance enough to miss the cutoff.

3. Causal inversions hidden in long sentences. When "X led to Y" becomes "Y led to X" inside a 30-word clause, most tokens stay identical, and cosine distance again changes only marginally.

Because every error sits within 0.010 of the boundary, simply lowering τ to a score such as 0.022 would raise recall on false claims without creating many false positives; but it would also erase any chance of labelling a quote true. The underlying limitation is that MiniLM embeddings are too coarse-grained for subtle factual reversals; swapping the encoder for a task-tuned cross-encoder or introducing a small margin band (Cannot Determine for 0.023–0.027) are straightforward ways to harden the system.

In short, with a hand-picked threshold CoVe-Kernel can spot blatant fabrications with high precision, but it remains brittle around fine-grained numeric or causal tweaks—exactly the corner cases that modern LLMs also find most challenging.

6 Discussion

Our evaluation of eight citation-verifying systems, including several advanced LLMs and one hybrid non-parametric method, reveals key trends about the strengths and limitations of current approaches to automated claim citation verification. The results demonstrate that while LLMs have made progress in factual reasoning, their ability to judge claim-evidence alignment consistently remains uneven, especially in adversarial or subtly perturbed contexts.

6.1 Performance vs. Prudence Tradeoff

A clear pattern emerges in the relationship between decisiveness and performance. Models like Gemini 1.5 Pro and Mistral-7B-Instruct, which issue definitive judgments with relatively low abstention rates (10.8% and 15.4%, respectively), achieve the highest overall accuracy and macro-F1 scores. In contrast, Claude 3 Opus and Perplexity 70B adopt a cautious stance, abstaining from over half the inputs, underperforming on both precision weighted and overall correctness. This emphasizes a central challenge in ethical LLM deployment: overly conservative models risk failing to flag misinformation, while confident ones may propagate falsehoods when it does not reflect factual correctness.

Furthermore, model size was not the primary determinant of performance. Despite having fewer parameters, Mistral-7B-Instruct outperformed several larger commercial systems, highlighting the value of instruction tuning and alignment strategies over raw scale. This suggests that accessible, open weight models, when carefully tuned, can achieve advanced performance in citation-sensitive tasks without requiring proprietary infrastructure.

6.2 Fine-Grained Factuality Remains Elusive

Both LLMs and the CoVe-Kernel method struggled with subtle perturbations, especially numeric

alterations and causal inversions. In contrast, the CoVe-Kernel system achieved 96.4% accuracy on its benchmark, with every error clustered near the decision threshold, revealing a sensitivity to edge cases. Such failure modes emphasize that vector distance, while capturing semantic similarity, is insufficient for ensuring factual equivalence. In practical terms, changing "42 million" to "41 million" or flipping cause-effect relationships produced only minor shifts in embedding space, small enough to evade detection by both LLMs and shallow similarity functions, highlighting a need for deeper analysis beyond word overlap in critical domains like journalism and legal review.

6.3 Ethical Implications and Design Considerations

Our findings carry several implications for the design and deployment of LLMs in citation-sensitive environments. First, models that over-rely on confidence or refuse to abstain when uncertain about data may contribute to hallucinated factuality, the illusion of truth created by authoritative tone and plausible structure. Second, the tendency of some models to abstain excessively raises the risk of ethical ambiguity, failing to identify misinformation when a judgment is expected.

The high performance of a relatively simple CoVe-Kernel baseline further raises questions about the interpretability and transparency of LLM outputs. Unlike most LLMs, which offer little insight into why a given citation was judged as accurate, the kernel-based method provides direct access to distance thresholds and can be calibrated to balance precision and recall. This suggests that hybrid systems, like our CoVE-Kernel system, may offer a more robust path forward for citation verification.

7 Conclusion

This study evaluated whether state-of-the-art LLMs can reliably distinguish true statements from minimally perturbed fabrications. We constructed a 1,000-item test set by pairing 500 verified AVeriTeC claims with single-edit counterparts, each manually validated to remove superficial cues. Seven zero-shot LLMs and a CoVe-Kernel baseline were assessed using label accuracy, macro-F1, and abstention rate.

Decisive models—Google Gemini 1.5 Pro (82.5 % accuracy) and Mistral-7B Instruct (87.0 % macro-F1)—consistently outperformed cautious systems such as Claude 3 Opus and Perplexity 70 B, which abstained on over half of the inputs and fell below 50 % overall accuracy. The CoVe-Kernel approach, relying on MiniLM embeddings with a single RBF cutoff, achieved 96.4 % accuracy, underscoring the competitiveness of simple, interpretable methods.

These results reveal a pronounced trade-off between decisiveness and restraint: lower abstention rates drive higher accuracy, whereas excessive hedging imposes substantial performance costs. Crucially, model scale alone does not determine success; instruction tuning and calibrated abstention thresholds are equally decisive.

Future work should (1) enhance small encoders or cross-encoders to detect subtle numeric and causal perturbations and (2) develop fully integrated pipelines that unify fine-grained citation ("sanitation"), systematic self-verification ("verification"), and atomic evaluation metrics such as FActScore. Such end-to-end frameworks promise to advance the reliability and transparency of LLM-based fact-verification systems.

8 References

References

[1] Menick, J.; Kadav, A.; Jaques, N.; Chen, M.; Petrov, M.; Hesse, C.; Clark, C. Teaching Language Models to Support Answers with Verified Quotes. *arXiv:2203.11147*, 2022.

[2] Dhuliawala, S.; Min, S.; Zhan, C.; Narayan-Chen, T.; Yasunaga, M.; McCann, B.; Prabhakaran, V. Self-Verification Improves Few-Shot Reasoning. *arXiv:2305.14251*, 2023.

[3] Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. 2023.

[4] Michel, G.; Epure, E. V.; Hennequin, R.; Cerisara, C. Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3. 2024.

[5] Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; Nakov, P. Factuality of Large Language Models: A Survey. In *Proceedings of EMNLP 2024*, 2024.

[6] Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv:2311.05232*, 2023.

[7] Zhang, Y.; Liu, S.; Qin, Z.; Wan, X.; Feng, Y. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv:2405.07437*, 2024.

[8] Barbero, A.; Carvalho, J.; Bode, N.; West, A.; Peterson, J. Robust Hallucination Detection in LLMs via Adaptive Token Selection. *arXiv:2504.07861*, 2025.

[9] Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: a Large-scale Dataset for Fact Extraction and Verification. In *EMNLP*, 2018.

[10] Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Cheng, J.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; Wang, Y.; Yang, L.; Wang, J.; Xie, X.; Zhang, Z.; Zhang, Y. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. *arXiv:2310.07521*, 2023.

[11] Chen, S.; Zhao, Y.; Zhang, J.; Chern, I.-C.; Gao, S.; Liu, P.; He, J. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *NeurIPS Workshops*, 2023.

[12] Zhang, H.; Anjum, S.; Fan, H.; Zheng, W.; Huang, Y.; Feng, Y. Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in LLMs. *arXiv:2503.16541*, 2025.

[13] Ma, H.; Xu, W.; Wei, Y.; Chen, L.; Wang, L.; Liu, Q.; Wu, S.; Wang, L. EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification. In *Findings of ACL*, pp. 9340–9349, 2024.

[14] Bazarova, A.; Yugay, A.; Shulga, A.; Ermilova, A.; Volodichev, A.; Polev, K.; Belikova, J.; Parchiev, R.; Simakov, D.; Savchenko, M.; Savchenko, A.; Barannikov, S.; Zaytsev, A. Hallucination Detection in LLMs with Topological Divergence on Attention Graphs. 2025.

[15] Peng, et al. Unanswerability Evaluation for Retrieval Augmented Generation. 2024.

[16] Fu, X.-Y.; Laskar, M. T. R.; Chen, C.; Tn, S. B. Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs. In *GEM Workshop at NeurIPS*, pp. 310–316, 2023.

[17] Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *TACL*, 5, 2017.

[18] Mausam; Schmitz, M.; Soderland, S.; Bart, R.; Etzioni, O. Open Language Learning for Information Extraction. In *EMNLP-CoNLL*, 2012.

[19] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In *ICLR*, 2018.

[20] Feher, D.; Khered, A.; Zhang, H.; Batista-Navarro, R.; Schlegel, V. Learning to Generate and Evaluate Fact-Checking Explanations with Transformers. *arXiv:2410.15669*, 2024.

[21] Pang, H.; Li, C.; Zhang, L.; Wang, S.; Zhang, X. Beyond Text: Fine-Grained Multi-Modal Fact Verification With Hypergraph Transformers. In *AAAI*, vol. 39, pp. 6389–639, 2025.

[22] Thakur, N.; Pradeep, R.; Upadhyay, S.; Campos, D.; Craswell, N.; Lin, J. Support Evaluation for the TREC20 4 RAG Track: Comparing Human versus LLM Judges. *arXiv:2504.15205*, 2025.

[23] Johnson, M.; Schuster, M.; Thorat, N.; Krikun, M.; Wu, Y.; Chen, Z.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; Dean, J. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *TACL*, 5, pp. 339–351, 2017. DOI:10.1162/tacl_a_00065

[24] Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; Etzioni, O. Open Information Extraction from the Web. In *IJCAI*, p .26 0–2676, 2007.

[25] Kriege, N. M.; Johansson, F. D.; Giscard, P.-L. A Survey on Graph Kernels. *arXiv:1903.11836*, 2019.

[26] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE TNNLS*, 32(1), pp. 4–24, 2021. DOI:10.1109/TNNLS.2020.2978386

[27] Nguyen, B.; Nguyen, L.; Dinh, D. Multi-level Community-awareness Graph Neural Networks for Neural Machine Translation. In *COLING*, pp. 5021–5028, 2022.

[28] Patel, M.; Anand, A. Factuality or Fiction? Benchmarking Modern LLMs on Ambiguous QA with Citations. *arXiv:2412.18051*, 2024.

[29] Tonmoy, S. M. I.; Zaman, S. M. M.; Jain, V.; Rani, A.; Rawte, V.; Chadha, A.; Das, A. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv:2401.01313*, 2024.

[30] Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; Nakov, P. Factuality of Large Language Models: A Survey. *arXiv:2402.02420*, 2024.