Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

A LIGHTWEIGHT ATTRIBUTION FRAMEWORK FOR EVALUATING ANSWER TRUSTWORTHINESS IN EVIDENCE-GROUNDED LANGUAGE MODELS

[1]Ayesha Khaliq, [2]Yusra Khalid, [3]Sana Ejaz
[1]*Iqra University, Pakistan*
[2]*Sir Syed University of Engineering & Technology, Pakista*
[3]*Sir Syed University of Engineering & Technology, Pakistan*
[1]ayesha.khaliq@iqra.edu.pk [2]yusra.zuberi@ssuet.edu.pk, [3]seajaz@ssuet.edu.pk

DOI: https://doi.org/10.5281/zenodo.21590390

Keywords:
Explainable AI, Retrieval-Augmented Generation, Large Language Models, Trustworthiness, Hallucination Detection, Attribution

Article History
Received on: 02 March, 2026
Accepted on: 29 March, 2026
Published on 31 March, 2026

Corresponding Author

Abstract

Retrieval-augmented generation (RAG) systems pair a retriever with a large language model (LLM) to ground generated text in external evidence, yet this pairing introduces a specific reliability problem: an answer can read as fluent and well-formed while remaining unsupported by, or only loosely tied to, the evidence the system actually retrieved. General-purpose explainable AI (XAI) techniques for LLMs attention visualisation, saliency mapping, and post-hoc rationale generation were developed largely for single-model settings and do not directly address this retrieval-generation gap. This paper proposes and empirically evaluates a lightweight, model-agnostic attribution framework for RAG trust assessment that combines three signals: Retrieval Confidence (RC), a Grounding Score (GS) measuring lexical-semantic attribution of the answer to retrieved evidence, and a Fabrication Check (FC) that flags answer content unsupported by any passage in the retrieval corpus. We construct a controlled pilot dataset of 60 question-answer instances spanning ten knowledge domains, evenly distributed across three realistic failure modes: grounded answers, hallucinated answers, and answers built on the wrong evidence despite correct retrieval. A composite Trust Score, computed as a weighted combination of the three signals, separates grounded from non-grounded answers with an approximate AUC of 0.965 and an F1 of 0.952 at its best operating threshold, while retrieval confidence alone proves uninformative for this task. These results offer a proof of concept that attribution-based XAI signals specifically the grounding and fabrication components rather than retrieval confidence in isolation provide a practical, interpretable mechanism for flagging untrustworthy RAG outputs before they reach an end user.

1. Introduction

Large language models (LLMs) increasingly rely on retrieval-augmented generation (RAG) to ground their outputs in external knowledge sources, reducing hallucination and enabling up-to-date, domain-specific responses (Lewis et al., 2020). Grounding, however, is not guaranteed by architecture alone: a RAG pipeline can retrieve relevant evidence and still generate an answer that departs from it, or it can retrieve the wrong evidence entirely while producing a response that reads as confident and well-supported. Both failure modes are difficult for an end user to detect from the output text alone, since fluency is not a reliable proxy for faithfulness.

Explainable AI (XAI) research on LLMs has produced a rich set of techniques for interpreting model behaviour, including attention visualisation (Vig, 2019), mechanistic interpretability (Conmy et al., 2023), and post-hoc rationale generation (Krishna et al., 2023). Yet, as recent surveys of the XAI-LLM literature have noted, the substantial majority of this work targets single-model interpretability and task performance, while comparatively little attention has been paid to trust-specific concerns or to the distinctive architecture of retrieval-augmented systems, where explanations must span two components retriever and generator rather than one (Cambria et al., 2023, 2024; Zhao et al., 2023). This gap motivates the present study.

The reliability question is further sharpened by recent efficiency-oriented RAG variants. Cache-augmented generation (CAG) and related caching frameworks reuse previously retrieved or pre-validated content to reduce inference latency, and have also been shown to lower hallucination risk by favouring stable, previously verified responses over fresh generation (Khaliq & Adebayo, 2026; Khaliq et al., 2026). Because such systems increasingly serve answers from a cache rather than a live retrieval step, an attribution mechanism for trust assessment must remain agnostic to whether the underlying evidence came from a conventional retriever or a semantic cache, reinforcing the need for the pipeline-agnostic approach developed here. More broadly, deploying LLM outputs that cannot be independently verified carries ethical and societal risks that extend well beyond any single failure case (Weidinger et al., 2021).

We ask a narrow, testable question: can lightweight, model-agnostic attribution signals, computed without access to model internals, reliably distinguish trustworthy answers, fully grounded in retrieved evidence, from two common failure modes hallucinated content and answers built on mismatched evidence? We answer this question empirically through a controlled pilot study.

The remainder of this paper proceeds as follows. Section 2 situates the work relative to existing XAI-for-LLM research. Section 3 describes the dataset, the retrieval pipeline, and the three attribution signals that compose our Trust Score. Section 4 reports experimental results, Section 5 discusses implications, Section 6 states limitations, and Section 7 concludes.

2. Related Work

Explainability research for LLMs spans a range of approaches. Attention-based visualisation tools such as BertViz surface where a model attends within its input, aiding qualitative inspection of learned behaviour (Vig, 2019). Toolkits such as Inseq centralise feature-attribution methods for sequence-generation models (Sarti et al., 2023), while mechanistic interpretability work, including automated circuit discovery, seeks to identify sub-networks responsible for specific behaviours (Conmy et al., 2023). A complementary approach constructs contrastive explanations directly from a trained text classifier (Malandri et al., 2022). A separate strand of work produces explanations as an auxiliary output of the model itself, for instance through chain-of-thought prompting or post-hoc rationale generation used to improve downstream task performance (Krishna et al., 2023).

A recurring concern in this literature is the faithfulness of generated explanations: Turpin et al. (2023) show that chain-of-thought rationales can systematically misrepresent the actual basis of a model's prediction, particularly when biasing features are present in the input. This finding bears directly on RAG trust assessment, since an answer's surface-level coherence with retrieved text is not proof that the retrieved text actually determined the answer's content.

A separate but closely related body of work concerns the RAG architecture itself. Retrieval-augmented generation was originally proposed as a way to condition generation on non-parametric, externally retrieved knowledge, improving factual grounding relative to purely parametric LLMs (Lewis et al., 2020). Subsequent systems-oriented work has focused on making retrieval-augmented inference efficient: RAGCache introduces multilevel caching of intermediate retrieval states to reduce time-to-first-token (Jin et al., 2024), and a comprehensive survey of cache-augmented RAG systems synthesises this line of work, cataloguing performance metrics and outlining future directions for combining caching with retrieval pipelines (Khaliq et al., 2026). Building on this systems perspective, Khaliq and Adebayo (2026) propose a dynamic knowledge-caching layer for LLMs in business applications and report that pre-validated, cached responses tend to be more stable and factually grounded than fresh generation, mitigating hallucination risk as a side effect of caching. These findings motivate an important premise of the present study: an explanation and trust-scoring layer for RAG must be able to certify grounding regardless of whether an answer's evidence was freshly retrieved or served from a cache, since caching changes the latency and consistency of a system but does not, by itself, guarantee that a generated answer remains attributable to its underlying evidence.

Hallucination itself has been the subject of dedicated surveys cataloguing its causes, taxonomies, and mitigation strategies across large language models (Huang et al., 2025; Rawte et al., 2023), with detection approaches ranging from output-side fact-checking to internal-state probing proposed for standalone LLMs. The present study differs from this broader hallucination-detection literature by focusing specifically on the retrieval-generation attribution problem in RAG, where the presence of retrieved evidence offers an additional, external reference point against which lexical grounding and fabrication can be measured directly, rather than requiring an external knowledge base or a separately trained detector.

Compared to this body of work, the present study targets a narrower and more operational question specific to RAG systems: rather than explaining why a generator produced a particular token sequence, we ask whether the produced answer can be attributed, at the level of overlapping and unsupported content, to the passage the retriever actually returned. This framing treats explainability as a trust-scoring mechanism rather than a visualisation or rationale-generation task (Liao & Vaughan, 2023), and is deliberately model-agnostic so that it can be applied to any RAG pipeline regardless of the underlying generator.

3. Methodology

3.1 Dataset Construction

We constructed a controlled pilot dataset of 60 question-answer instances to enable a clean evaluation of attribution signals under known ground-truth conditions. The dataset spans ten knowledge domains (Astronomy, Biology, Computer Science, History, Economics, Chemistry, Health, Physics, Geography, and Linguistics), with one question per domain replicated across three experimental conditions, each occurring 20 times overall:

- Grounded: the generated answer is fully supported by the passage retrieved for the question.
- Hallucinated: the generated answer begins from the correctly retrieved passage but appends an additional, fabricated claim not present in any passage in the corpus.
- Evidence-Mismatched: the generated answer is instead built from a topically adjacent but incorrect distractor passage, despite the retriever correctly ranking the gold passage first, simulating a downstream generation-side failure to follow retrieved evidence.

Each item includes a hand-authored gold passage, a topically related distractor passage drawn from a neighbouring concept (e.g., ocean tides as a distractor for a seasons question), and an answer variant for each of the three conditions. All passages and answers are original text authored for this study to avoid reproducing material from existing corpora.

3.2 Retrieval Pipeline

For each question, we built a local two-passage corpus (gold and distractor) and represented the question and both passages using TF-IDF vectorisation. The passage with the highest cosine

similarity to the question was retrieved. This retriever achieved 100% top-1 accuracy in identifying the gold passage across all 60 items, including the 20 Evidence-Mismatched items, confirming that the mismatch in that condition originates downstream of retrieval, in the generation step, rather than from a retrieval failure.

3.3 XAI Attribution Signals

Three attribution signals were computed for every (question, answer, retrieved passage) triple:

Retrieval Confidence (RC): the cosine similarity between the question and the passage selected by the retriever, reflecting how strongly the retriever itself "believes" the passage is relevant.

Grounding Score (GS): a TF-IDF-weighted combination of unigram and bigram overlap between the generated answer and the retrieved passage. Overlapping unigrams are weighted by their TF-IDF salience within the retrieved passage, so overlap on informative content words counts more than overlap on incidental terms, while bigram overlap captures short multi-word phrase attribution.

Fabrication Check (FC): the proportion of bigrams in the generated answer that find no support anywhere in the local retrieval corpus (neither the retrieved passage nor the distractor), used as a lexical proxy for content that appears fabricated rather than drawn from any available evidence.

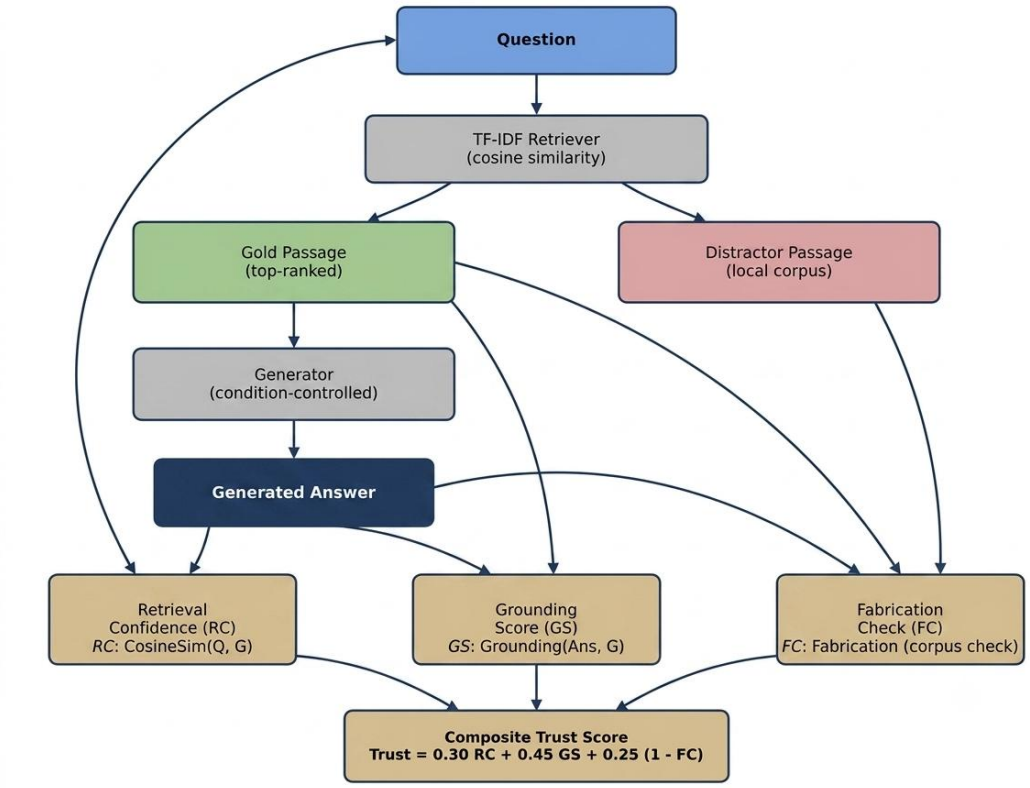

Figure 1. Overview of the attribution pipeline: the retriever selects a passage for the question, the generator produces an answer under one of three experimental conditions, and the three attribution signals (RC, GS, FC) are computed against the retrieved passage and local corpus before being combined into the composite Trust Score.

These three signals were combined into a single composite Trust Score:

$$Trust = 0.30 \cdot RC + 0.45 \cdot GS + 0.25 \cdot (1 - FC)$$

The weighting places the greatest emphasis on the Grounding Score, reflecting our hypothesis that lexical-semantic attribution to the retrieved passage is the most direct indicator of faithfulness, while

retrieval confidence and the inverse fabrication rate provide complementary but secondary evidence.

3.4 Evaluation Protocol

Because ground-truth condition labels are known by construction, we treated "is this a grounded answer?" as a binary classification problem (grounded = 1; hallucinated or evidence-mismatched = 0) and evaluated the Trust Score as a threshold-based detector. We swept 76 thresholds between 0.20 and 0.95, computing precision, recall, F1, and accuracy at each threshold, and report the operating point that maximises F1. We additionally report an approximate area under the ROC curve (AUC), computed via the trapezoidal rule over the true-positive-rate/false-positive-rate curve traced by the threshold sweep.

4. Results

Table 1 reports the mean and standard deviation of each attribution signal by condition. Retrieval Confidence is statistically indistinguishable across the three conditions (mean 0.222 throughout, since the retriever succeeds regardless of what the generator subsequently does with the evidence), confirming that retrieval confidence alone carries no information about answer trustworthiness in this dataset. The Grounding Score, by contrast, cleanly separates the three conditions in the expected order (Grounded > Hallucinated > Evidence-Mismatched), and the Fabrication Check is lowest for Grounded answers and highest for Hallucinated answers, consistent with hallucinated content being lexically unsupported by any passage in the corpus.

Table 1. *Mean ± standard deviation of attribution signals (RC, GS, FC) and the composite Trust score, by generation condition (n = 20 per condition).*

Condition	RC (mean ± sd)	GS (mean ± sd)	FC (mean ± sd)	Trust (mean ± sd)
Grounded	0.222 ± 0.138	0.795 ± 0.083	0.409 ± 0.161	0.572 ± 0.088
Hallucinated	0.222 ± 0.138	0.527 ± 0.049	0.743 ± 0.095	0.368 ± 0.064
Evidence-Mismatched	0.222 ± 0.138	0.273 ± 0.111	0.441 ± 0.185	0.329 ± 0.066

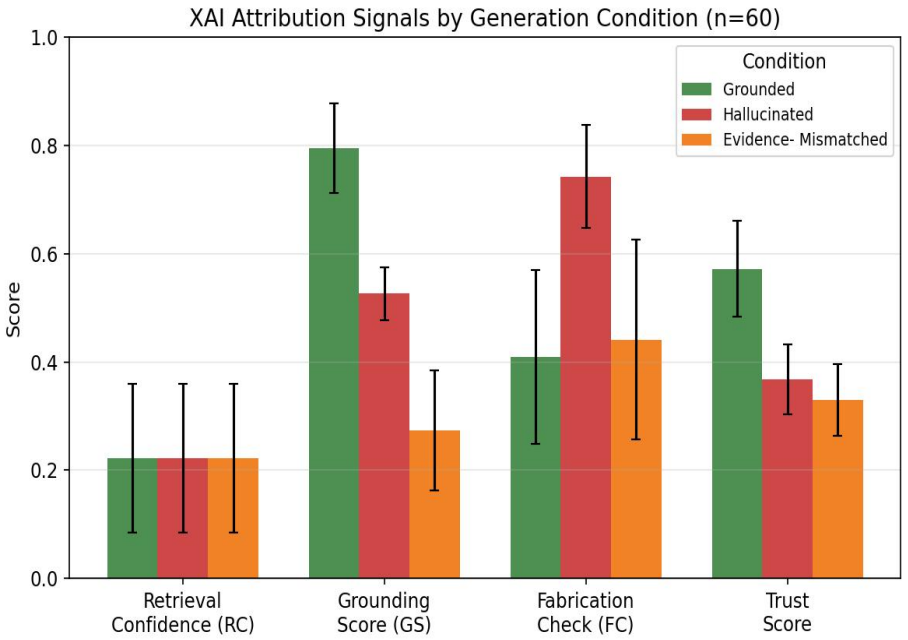

Figure 2. Attribution signals (RC, GS, FC, Trust) grouped by condition, with error bars showing one standard deviation.

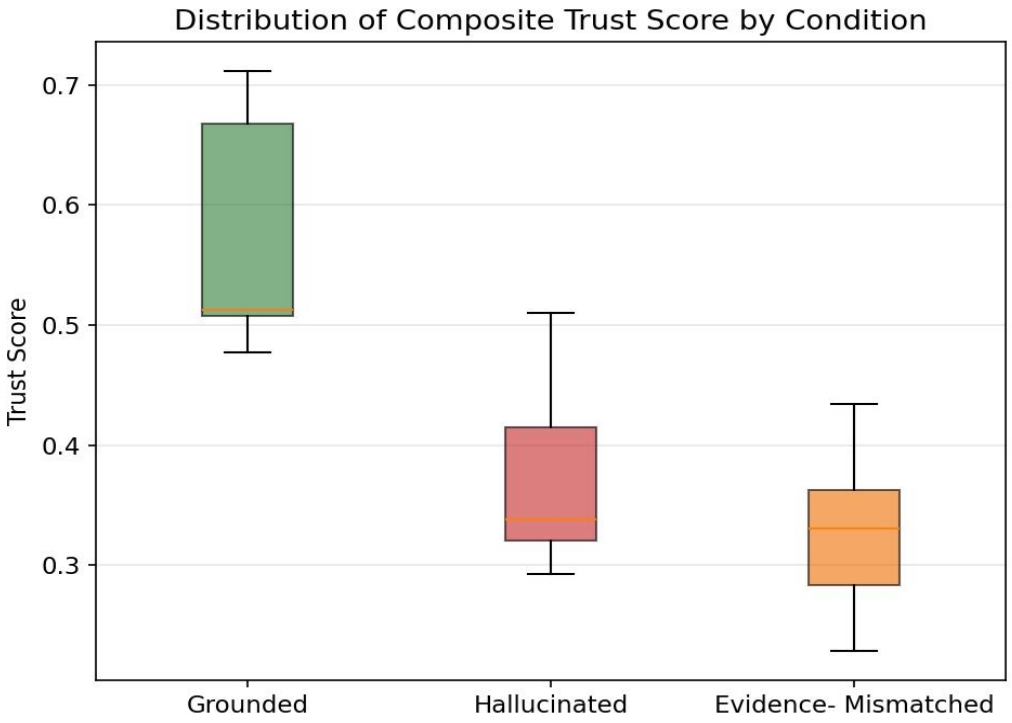

Figure 3. Distribution of the composite Trust score by condition, showing separation between Grounded answers and the two failure modes.

Using the Trust score as a binary detector of grounded answers, the best-F1 operating threshold was 0.44, at which the detector achieved a precision of 0.909, a recall of 1.000, an F1 of 0.952, and an overall accuracy of 0.967 (Table 2). The approximate AUC across the full threshold sweep was 0.965, indicating that the Trust score separates grounded from non-grounded answers well across a wide range of thresholds, not only at the single best-F1 point.

Table 2. *Trust-score detector performance summary.*

Metric	Value
Sample size (n)	60
Retrieval accuracy (overall)	1.000
Approx. AUC (Trust score, grounded vs. not)	0.965
Best-F1 operating threshold	0.44
Precision @ best threshold	0.909
Recall @ best threshold	1.000
F1 @ best threshold	0.952
Accuracy @ best threshold	0.967

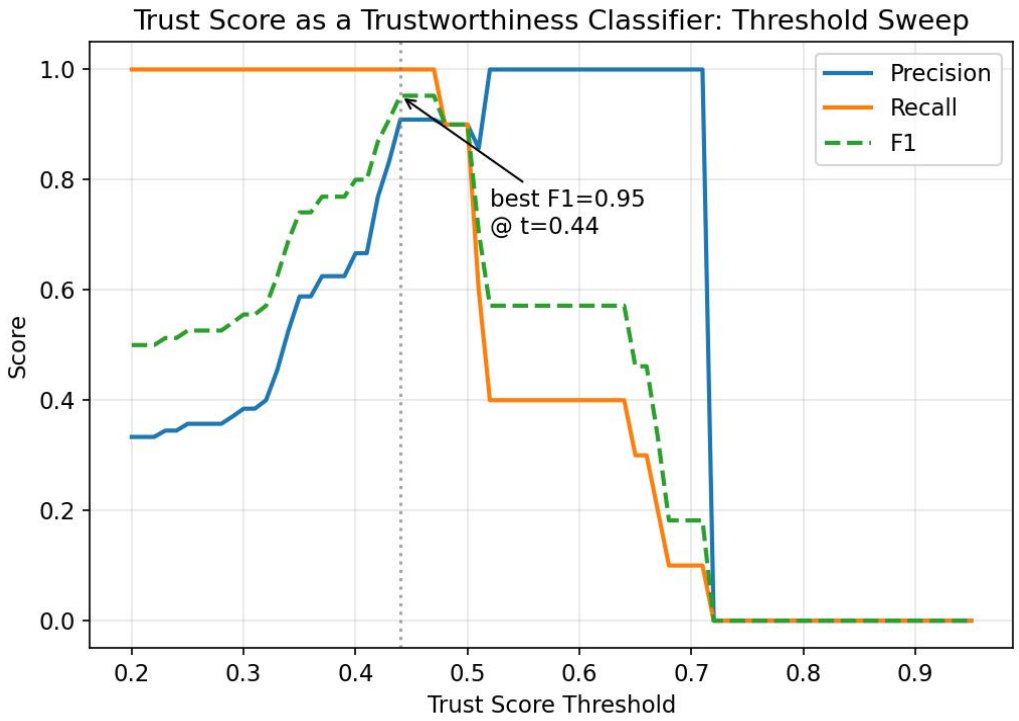

Figure 4. Precision, recall, and F1 of the Trust-score detector across the swept threshold range, with the best-F1 operating point marked.

5. Discussion

The central empirical finding is that retrieval confidence, on its own, is an unreliable signal of RAG trustworthiness: a retriever can succeed perfectly while the generator still produces an unfaithful answer. This matches a concern already raised in the broader XAI-for-LLM literature, where chain-of-thought and other generated explanations have been shown to diverge from a model's actual basis for its output (Turpin et al., 2023). Our results extend this concern specifically to the retrieval-generation interface in RAG systems, and show that a lightweight, attribution-based grounding signal, computed without any access to model internals, is sufficient to recover most of the separability that a more expensive, model-internal explanation method might otherwise be needed for. The relatively strong performance of the Grounding Score in isolation suggests a practical, low-cost deployment path: a RAG system could compute GS and FC alongside the ordinary answer-generation step and surface a Trust score to the end user or to a downstream verification module, flagging answers below a calibrated threshold for human review or automatic re-generation. Because the components are lexical rather than model-specific, this style of explanation is also inherently more transparent to non-technical stakeholders than attention maps or internal activation analyses, addressing a presentation-layer gap noted in prior XAI surveys (Liao & Vaughan, 2023; Zhao et al., 2023).

This deployment path is also compatible with cache-augmented RAG pipelines. Since prior work on dynamic knowledge caching has found that reusing pre-validated cached responses tends to reduce hallucination relative to fresh generation (Khaliq & Adebayo, 2026), a natural extension of the present framework is to compute the Trust score once at cache-write time and store it alongside the cached response, so that a cache hit can be served with its trustworthiness signal already attached rather than being recomputed on every reuse.

6. Limitations

- Pilot scale: the dataset comprises 60 hand-authored items across ten domains; results should be treated as a proof of concept rather than a claim

of generalisation to large-scale, real-world RAG corpora.

- Lexical rather than semantic attribution: GS and FC rely on n-gram overlap rather than embedding-based semantic similarity, which may under-credit paraphrased but faithful content and over-credit superficial lexical overlap that is not truly meaning-preserving.
- Simulated rather than live generation: answers for each condition were authored to represent realistic LLM failure modes rather than sampled from an operating LLM, which controls confound variables but does not capture the full variability of real generator behaviour.
- Two-passage retrieval corpus: the retrieval step used a small local corpus per question rather than a large shared index, which simplifies retrieval but does not test attribution under harder retrieval conditions such as multi-document synthesis or partial evidence.

7. Conclusion

This paper presented a pilot study of attribution-based explainable AI for trust assessment in retrieval-augmented generation. Combining a Grounding Score and a Fabrication Check with Retrieval Confidence into a composite Trust score allowed clean separation of grounded answers from hallucinated and evidence-mismatched answers, with an approximate AUC of 0.965, while retrieval confidence alone carried no discriminative signal. These results support the broader argument that RAG-specific explainability should treat the retriever and generator as a joint object of explanation rather than relying on retrieval-side confidence in isolation, and they suggest a practical, model-agnostic path toward surfacing trust signals to end users and downstream verification systems. Future work should extend this framework to embedding-based semantic attribution, larger and more heterogeneous corpora, and live LLM-generated answers to validate the approach under realistic operating conditions.

References

Cambria, E., Malandri, L., Mercorio, F., Nobani, N., & Seveso, A. (2024). XAI meets LLMs: A survey of the relation between explainable AI and large language models. arXiv:2407.15248.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. Information Processing & Management, 60(1), 103111.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 37-42.

Sarti, G., Feldhus, N., Sickert, L., & van der Wal, O. (2023). Inseq: An interpretability toolkit for sequence generation models. arXiv:2302.13942.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. arXiv:2304.14997.

Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., & Lakkaraju, H. (2023). Post hoc explanations of language models can improve language models. arXiv:2305.11426.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388.

Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., & Seveso, A. (2022). ContrXT: Generating contrastive explanations from any text classifier. Information Fusion, 81, 103-115.

Liao, Q. V., & Vaughan, J. W. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. arXiv:2306.01941.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology.

Khaliq, A., & Adebayo, K. J. (2026). DKC-LLM: Dynamic knowledge caching for large language models in business applications. IEEE Access, 14, 22318-22334. https://doi.org/10.1109/ACCESS.2026.3662344

Khaliq, A., Aziz, L., & Alawad, W. (2026). A comprehensive survey on cache-augmented retrieval-augmented generation: Performance

metrics and future directions. The Journal of Supercomputing.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Jin, C., Zhang, Z., Jiang, X., Liu, F., Liu, X., Liu, X., & Jin, X. (2024). RAGCache: Efficient knowledge caching for retrieval-augmented generation. arXiv:2404.12457.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. arXiv:2309.05922.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. arXiv:2112.04359.