Available Online: 21 July 2026 DOI: 10.54254/2977-3903/2026.35471

Robust and explainable Retrieval-Augmented Generation under retrieval noise

Shirui Chen

School of Science and Technology, James Cook University, Singapore, Singapore

chenshirui277443@163.com

Abstract. Retrieval-Augmented Generation (RAG) improves language-model answers by retrieving external evidence before generation. However, its reliability depends on the retrieved context. In real settings, passages can be irrelevant, incomplete, conflicting, or poorly ordered. These problems may reduce accuracy and explainability. This study tests how retrieval noise affects RAG and whether reranking, citation-aware generation, and lightweight verification can improve system behaviour. A controlled experiment was conducted on a small HotpotQA subset using BM25, Sentence-BERT, and FAISS. Four systems were compared: vanilla RAG, reranking-only, citation-only, and a full enhanced system. Results show that reranking achieved the highest average noisy F1, but the gain over vanilla RAG was small. The full enhanced system achieved better faithfulness, groundedness, and citation precision, but did not improve average noisy F1. This suggests that robust and explainable RAG is a multi-objective problem.

Keywords: Retrieval-Augmented Generation, retrieval noise, robustness, explainability, evidence reranking

1. Introduction

Retrieval-Augmented Generation (RAG) combines external retrieval with language-model generation to improve factuality and to ground answers in evidence. A RAG system first retrieves passages and then uses them to generate an answer. However, retrieval is useful only when the passages are relevant, complete, and clearly organised.

This is important because real retrieval is often noisy. Retrieved context may include irrelevant passages, incomplete evidence, conflicting information, or useful evidence in a weak position. Previous studies show that RAG should be evaluated based on answer accuracy, faithfulness, groundedness, and failure behaviour [1-3]. Other work also shows that retrieval can help or hurt depending on evidence quality and order [4-6].

This study addresses this issue through a small-scale controlled experiment on retrieval noise. It compares four RAG systems under clean and noisy settings. The analysis focuses on three conditions: the effect of retrieval noise on answer quality, the relative difficulty of different noise types, and the extent to which reranking, citation-aware generation, and lightweight verification are associated with robustness and explainability.

The experiment uses a HotpotQA distractor validation subset and a hybrid retriever combining BM25 and dense retrieval. The main contribution is a controlled ablation study that can provide a reference for future

RAG design under noisy retrieval conditions. It also helps identify likely risks in evidence use and offers practical suggestions for more reliable and explainable RAG development.

2. Literature review

Recent RAG evaluation has moved beyond answer correctness. RAGAS measures answer relevance, faithfulness, and context quality [1]. ARES focuses on answer relevance, faithfulness, and context relevance [2]. RAGChecker separates retrieval-side and generation-side failures [3]. These studies show that a RAG system should be evaluated based on both the final answer and its evidence use.

RAG robustness under imperfect retrieval is a growing problem. Retrieval augmentation may help or hurt, depending on the evidence retrieved [4]. RARE studies RAG under perturbed retrieval conditions, while other work shows that document order, retrieval errors, and evidence quality affect final performance [5, 6]. Therefore, retrieval noise is an important factor in answer quality.

Evidence refinement and explainability are also important in RAG research. Corrective Retrieval Augmented Generation argues that raw top-k retrieval is often not enough and that systems should check retrieval quality [7]. SEER shows that better evidence extraction can improve passage selection [8]. Groundedness research shows that an answer may look correct but still be weakly supported at the claim level [9]. Citation-aware generation can help evidence attribution, but citations should not be treated as proof that the answer is correct.

Another related issue is position. Hsieh's study shows that language models can be sensitive to where relevant information appears in long contexts [10]. This means that even if the same evidence is retrieved, the answer may change when the passage order changes.

Previous studies provide useful methods for evaluation, retrieval correction, groundedness analysis, and robustness testing. However, a small-scale controlled ablation is still useful because it can compare different enhancement strategies under the same retrieval setting. This study compares a vanilla baseline, a reranking-only system, a citation-only system, and a full enhanced system. This design separates reranking from citation-aware prompting and tests whether the full system improves robustness, explainability, or both.

3. Methodology

3.1. Experimental design

This study uses a controlled experiment to test the effect of retrieval noise on RAG robustness and explainability. The system has two main stages: retrieval and generation. The hybrid retriever combines BM25, Sentence-BERT embeddings, and FAISS vector search. The dataset is a small subset of the HotpotQA distractor validation set, containing 50 questions. For each question, the retriever returns five passages, and the generator uses the google/flan-t5-small model.

Four RAG configurations are compared, vanilla_rag uses the noisy retrieved passages with a standard prompt; enhanced_rerank_only refines and reranks the evidence after noise is added, but it keeps the same prompt. enhanced_citation_only uses the same noisy passages as the baseline but applies a citation-aware prompt; enhanced_full combines reranking, citation-aware generation, and a simple verification step.

To simulate imperfect retrieval, noise is added to the retrieved passages. The experiment uses four noise types: irrelevant noise, partially relevant noise, conflicting noise, and ordering noise. The noise ratios are set to 0.2, 0.4, and 0.6, while the clean condition has a ratio of 0.0. Therefore, each system is tested under 13 conditions.

$$\text{Noise Ratio} = \frac{\text{Number of Noisy Passages}}{\text{Total Retrieved Passages}} \quad (1)$$

Answer quality is measured by Exact Match and F1 score. Retrieval performance is measured by retrieval recall at k. Explainability is evaluated through faithfulness, groundedness, citation precision, citation presence, and unsupported-claim flags [1-3].

3.2. Experimental procedure

The experiment has five stages. First, the HotpotQA subset is prepared so each question has a gold answer and supporting context. Second, the hybrid retriever collects the clean top-k context. Third, synthetic noise is added. For irrelevant, partially relevant, and conflicting noise, some of the original passages are replaced or mixed with noisy passages. For ordering noise, only the passage order is changed.

Fourth, the four RAG systems are tested under the same clean and noisy conditions. Fifth, the results are summarized by system, noise type, and noise ratio. Since there are four systems and 13 conditions per system, the summary contains 52 data rows. With 50 samples, the raw output contains 2,600 rows.

$$\text{Robustness Drop F1} = \frac{\text{F1}_{\text{clean}} - \text{F1}_{\text{noisy}}}{\text{F1}_{\text{clean}}} \quad (2)$$

$$\text{Citation Precision} = \frac{\text{Number of Supported Cited Claims}}{\text{Total Number of Cited Claims}} \quad (3)$$

The study also includes qualitative case analysis. The case is used as illustrative evidence, not representative evidence. It shows how noise can change one answer and how reranking or citation support may correct some failures [1, 3, 9].

4. Results

4.1. Evaluation overview

The evaluation includes 52 aggregated conditions across four RAG systems, and each condition uses 50 examples. Therefore, the analysis is based on system-generated experimental outputs rather than expected or manually created values.

Three main patterns appear. First, retrieval noise reduced answer quality in most settings. Second, reranking gave only a small improvement in noisy F1. Third, citation-aware generation improved explainability more clearly than answer quality.

4.2. Answer quality under retrieval noise

As shown in Figure 1, mean F1 changed across systems and usually decreased under noise conditions. In the clean setting, enhanced_rerank_only had the highest mean F1 of 0.446, followed by enhanced_full at 0.444. vanilla_rag reached 0.384, while enhanced_citation_only had the lowest clean F1 of 0.325.

Under noisy conditions, enhanced_rerank_only achieved the highest average noisy F1 at 0.243, followed very closely by vanilla_rag at 0.239. The difference was only 0.004, so the robustness benefit from reranking was modest. enhanced_citation_only reached 0.208, and enhanced_full reached 0.215, both lower than the vanilla baseline in noisy answer quality.

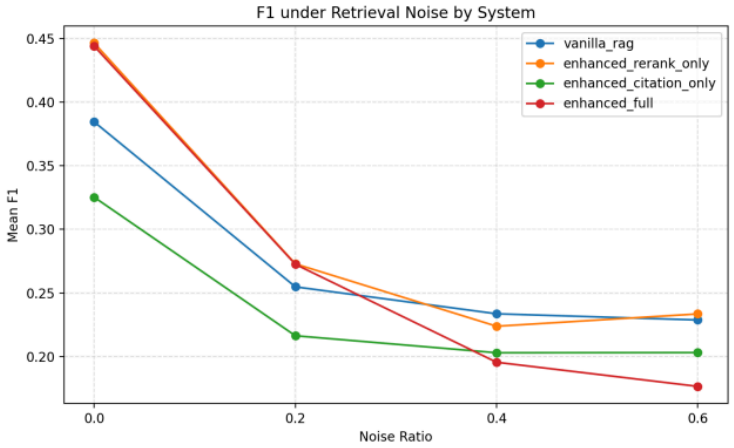

Figure 1. Mean F1 across noise ratios for the four RAG configurations

Table 1 reports the average noisy performance across all non-clean conditions.

Table 1. Aggregate noisy performance by system

System	Average noisy F1	Average robustness drop	Faithfulness	Groundedness	Citation precision
enhanced_rerank_only	0.243	0.455	0.390	0.635	0.000
vanilla_rag	0.239	0.378	0.399	0.642	0.000
enhanced_full	0.215	0.516	0.901	0.737	0.671
enhanced_citation_only	0.208	0.362	0.750	0.616	0.550

4.3. Robustness across noise types

Figure 2 shows that conflicting noise was the most damaging condition for vanilla_rag, with an average robustness drop of 0.665. This is reasonable because conflicting evidence gives misleading information that directly competes with the correct answer.

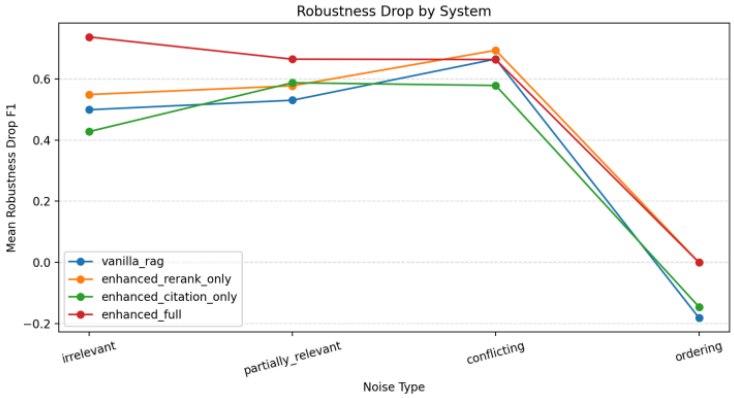

Figure 2. Mean robustness drop in F1 by noise type and system

Table 2 shows F1 by noise type. Under conflicting noise, all systems had low F1 scores, from 0.129 for vanilla_rag to 0.150 for enhanced_full. Under irrelevant noise, enhanced_rerank_only achieved 0.201, slightly above vanilla_rag at 0.193. Under partially relevant noise, it also performed slightly better, with 0.189 compared with 0.181. This suggests that reranking helps when useful evidence is present but needs to be prioritised.

Table 2. Average F1 by noise type

Noise type	vanilla_rag	enhanced_rerank_only	enhanced_citation_only	enhanced_full
conflicting	0.129	0.137	0.137	0.150
irrelevant	0.193	0.201	0.186	0.117
partially_relevant	0.181	0.189	0.134	0.149
ordering	0.454	0.446	0.372	0.444

4.4. Ordering noise effect

The ordering-noise results were unexpected. Instead of reducing F1, ordering noise was linked with higher average F1 for vanilla_rag and enhanced_citation_only. For vanilla_rag, F1 increased from 0.384 in the clean setting to 0.454 under ordering noise. The best performance under noise was also vanilla_rag under ordering noise at a ratio of 0.4, where mean F1 reached 0.465.

This should not be seen as a general benefit of ordering noise. A more likely explanation is a positional effect. Since FLAN-T5-small is a relatively small generator, reordering may accidentally move useful evidence into positions that are easier for the model to use [10].

4.5. Explainability metrics

Figure 3 shows the explainability metrics. enhanced_full achieved the strongest results, with average noisy faithfulness of 0.901, groundedness of 0.737, and citation precision of 0.671. It also had an unsupported-claim rate of 0.000. This suggests stronger evidence linkage when reranking, citation-aware generation, and verification are combined.

The citation-only system also improved explainability compared with the non-citation systems. It had an average noisy citation precision of 0.550 and a citation presence rate of 1.000, but its noisy F1 was still lower than vanilla_rag. This shows that citation-aware generation can improve evidence attribution without necessarily improving answer correctness.

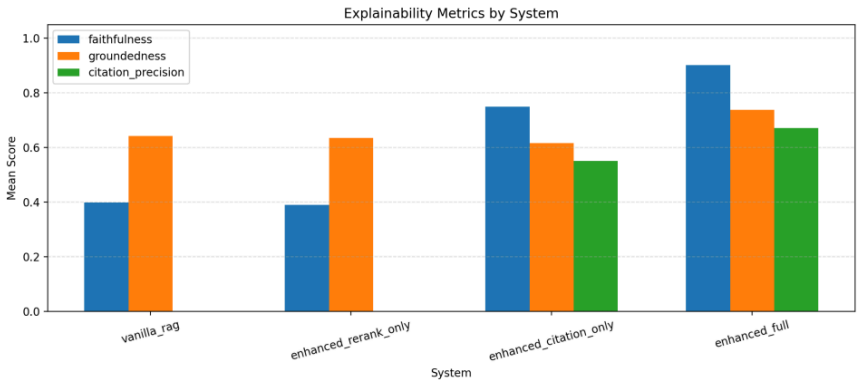

Figure 3. Explainability-related metrics across the four RAG configurations

4.6. Case-level evidence

The case analysis shows how retrieval noise can affect individual answers. In one conflicting-noise example, the question asked which band had more members, Test Icicles or X Ambassadors. The gold answer was X Ambassadors. vanilla_rag predicted Test Icicles and received an F1 of 0.0, while the enhanced_full system predicted X Ambassadors with citations and achieved an F1 of 1.0.

This case is illustrative, not representative. It shows one possible correction, but it does not prove that enhanced_full is better overall. The aggregate results do not show a general robustness advantage for enhanced_full.

5. Discussion

The results show a trade-off between robustness and explainability. The reranking-only system achieved the best average noisy F1, but the improvement over vanilla_rag was small. In contrast, enhanced_full achieved the strongest explainability scores but did not improve noisy F1 over the baseline. Therefore, the enhanced design improves explainability more clearly than robustness.

This matters because citation-aware generation and lightweight verification can make outputs more transparent, but they do not always make the answers more accurate.

The reranking-only results suggest that evidence refinement can give a modest improvement under noisy retrieval. However, the gain is small, and reranking alone cannot fully solve retrieval noise. With conflicting passages, a reranker may still fail to separate correct evidence from misleading evidence. This agrees with work on corrective retrieval and evidence extraction [5, 7, 8].

The citation-only system improved citation behaviour but reduced average noisy F1. This suggests that citation-aware prompting may make generation harder while improving explainability. The full enhanced system achieved better faithfulness, groundedness, and citation precision than the other systems, but stronger evidence attribution should not be treated as the same as stronger answer correctness.

The ordering noise results suggest a possible positional effect, especially for vanilla_rag and enhanced_citation_only. Robustness depends on which passages are retrieved and how they are arranged in the prompt [10]. The experiment also has limitations: each condition uses only 50 examples, the results report means without confidence intervals or significance tests, the noise is synthetic, and the generator is a small instruction-tuned model. Some explainability metrics are partly structural, so they should be interpreted together with answer quality.

The findings suggest that robust and explainable RAG should be treated as a multi-objective problem. Reranking can slightly improve answer quality under noisy conditions, while citation-aware generation and verification can improve transparency and groundedness. However, combining these components does not guarantee better robustness.

6. Conclusion

This study examines how retrieval noise affects answer quality, robustness, and explainability in RAG systems. The experiment compared vanilla RAG, reranking-only, citation-only, and a full enhanced system under clean and noisy retrieval settings. The results show that retrieval noise is an important failure factor. Conflicting noise caused the clearest drop because it introduced misleading evidence that competed with the correct answer.

The findings also show that improvements were not the same across objectives. The reranking-only system achieved the highest average noisy F1, but its gain over vanilla RAG was very small. Citation-only enhancement improved citation behaviour but reduced answer quality when used alone. The full system achieved the strongest explainability results, but it did not outperform vanilla RAG in average noisy F1. These results do not prove that the full enhanced system is generally better. Therefore, robustness and explainability are related but different goals.

This study has several limitations. The sample size was only 50 examples per condition, and the results do not include confidence intervals or significance tests. The retrieval noise was synthetic, and the generator was relatively small. Future work should test larger datasets, stronger generators, more realistic retrieval failures, and more rigorous statistical evaluation. Future work could also study citation strategies that keep explainability while reducing the performance cost of citation prompting.

References

[1] Shahul, E., et al. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

[2] Saad-Falcon, J., et al. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

[3] Ru, D., et al. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 37*.

[4] Maekawa, S., et al. (2024). Retrieval helps or hurts? A deeper dive into the efficacy of retrieval augmentation to language models. In *NAACL 2024*.

[5] Zeng, Y., et al. (2025). RARE: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *arXiv*. https://arxiv.org/abs/xxxx.xxxxxx

[6] Cao, S., et al. (2025). Evaluating the retrieval robustness of large language models. *arXiv*. https://arxiv.org/abs/xxxx.xxxxxx

[7] Yan, S.-Q., et al. (2024). Corrective retrieval augmented generation. In *The Twelfth International Conference on Learning Representations*.

[8] Zhao, X., et al. (2024). SEER: Self-aligned evidence extraction for retrieval-augmented generation. In *EMNLP 2024*.

[9] Stolfo, A. (2024). Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of NAACL 2024*.

[10] Hsieh, C.-Y., et al. (2024). Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of ACL 2024*.