

Comparative Analysis of Confidence-Aware and Verification-Based Hallucination Reduction Techniques in Retrieval-Augmented Generation Systems

Khushboo Buwade¹ and Mitali Bedmutha

¹Specialist Programmer, Infosys Limited, Pune, India; khushbuwade183@gmail.com

ZSpecialist Programmer, Infosys Limited, Pune, India;

mitalibedmutha2002@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) grounds language-model responses in externally retrieved evidence, yet weak retrieval or unsupported generation can still reduce reliability. This pilot study compares two lightweight safeguards: retrieval-confidence gating and post-generation evidence verification. Four configurations were evaluated on the same 30-question benchmark: Basic RAG, Confidence-Aware RAG, Verification-Based RAG, and a Combined approach. The benchmark used 30 short passages, 25 answerable questions, and five deliberately unanswerable questions. Retrieval employed normalized all-MiniLM-L6-v2 embeddings with FAISS top-3 search; generation and verification used Groq Compound Mini. All 120 runs completed without API errors. Basic and Verification-Based RAG achieved 30/30 preliminary correct outcomes. Confidence-Aware and Combined RAG achieved 29/30 because one answerable query received a similarity score of 0.261 and was rejected by the fixed 0.55 threshold. No unsupported final answer was observed. Mean response times were 0.50 s, 0.40 s, 1.29 s, and 0.95 s, respectively. The results reveal a practical trade-off: confidence gating can reduce model calls and encourage safe refusal, but an uncalibrated threshold may reduce answer coverage, while verification preserves coverage at additional latency.

Key Words: Retrieval-Augmented Generation, evidence verification, retrieval confidence, abstention, hallucination, FAISS.

1. INTRODUCTION

An enterprise question-answering system must do more than produce fluent text. It must know when the available material supports an answer and when the safer action is to refuse. Large language models can generate statements that sound credible even when evidence is missing. RAG reduces this exposure by placing retrieved passages in the prompt, yet two uncertainties remain: the retriever may return a poor match, and the generator may add content not present in that match [1]-[3].

This work studies two checks that can be attached to an existing RAG pipeline without training a new model. Retrieval confidence is represented by the highest normalized similarity score produced by FAISS. Answer verification is implemented as a second prompt that compares the draft with the retrieved passages. These checks are evaluated separately and together so that their effects on correctness, refusal behavior, and response time can be observed.

The paper is an experimental comparison, not a proposal for a new learning algorithm. Its practical contribution is a small reproducible test that illustrates where simple safeguards help and where they create new failure modes. The main question is whether confidence gating and evidence verification improve reliability without causing unacceptable loss of coverage or latency.

2. BACKGROUND AND RELATED STUDIES

The original RAG formulation combined a sequence generator with retrieved non-parametric memory for knowledge-intensive tasks [1]. Dense Passage Retrieval demonstrated that question and passage

vectors can support effective open-domain evidence retrieval [8]. Similarity search libraries such as FAISS make this pattern practical for small and large vector collections [9], while Sentence-BERT-style encoders provide compact semantic representations [10].

Hallucination research distinguishes unsupported content from content that directly contradicts a source. Surveys by Ji et al. and Huang et al. describe causes, evaluation methods, and mitigation strategies across modern generation systems [2], [3]. Their analyses support an evidence-centred evaluation in which refusal is treated differently from an incorrect assertion.

Several RAG extensions examine the quality of retrieved evidence or the generated answer. Self-RAG introduces learned reflection behaviour for retrieval and critique [4]. Corrective RAG evaluates document quality and initiates corrective retrieval actions [5]. Chain-of-Verification asks a model to inspect and revise an initial response [6]. RAGAS separates retrieval quality, faithfulness, and answer relevance during evaluation [7]. The present implementation borrows these broad ideas but intentionally uses a simpler threshold and a prompt-based checker so that the study can be reproduced with limited resources.

3. EXPERIMENT DESIGN

3.1 Four configurations

Basic RAG retrieves three passages and asks the model to answer only from them. Confidence-Aware RAG performs the same retrieval but returns `INSUFFICIENT_EVIDENCE` when the best score is below 0.55. Verification-Based RAG always creates a draft, then asks a second model call to label the draft as supported, partly supported, or unsupported and to revise it when required. Combined RAG first applies the confidence gate and invokes verification only for accepted retrievals.

3.2 Retrieval and generation

Each corpus passage and question was encoded with all-MiniLM-L6-v2. Vectors were normalized and searched through FAISS IndexFlatIP, so the returned inner product behaves as cosine similarity. The top three passages were inserted into an evidence-only prompt. Groq Compound Mini generated the answer at temperature 0.01. For the verification methods, the same service received the question, evidence, and draft and returned a structured judgment. A fixed four-second delay between requests reduced API-rate pressure.

3.3 Benchmark

The local collection contains 30 short passages from science, computing, and AI. The evaluation set also contains 30 questions. Twenty-five questions have direct support in the collection; five were deliberately selected so that no supporting passage existed. The absent questions test whether a method invents an answer or reports insufficient evidence. Every question was processed by each configuration, giving 120 recorded outcomes.

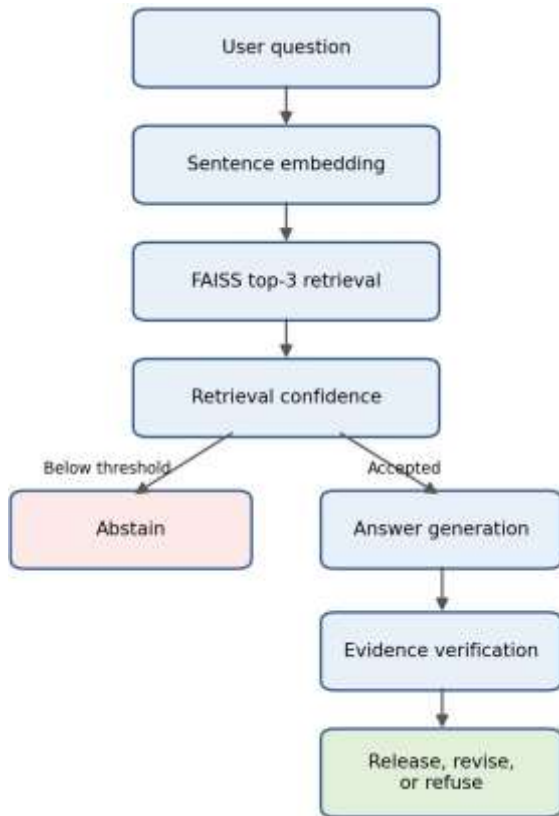


Fig -1: Execution path used for confidence gating and evidence verification.

4. MEASURES

Preliminary accuracy was computed from reference answers and required answer terms. A response was marked as a hallucination when it asserted an unsupported answer for an absent question or produced an incorrect supported answer instead of abstaining. The abstention rate is the proportion of outputs equal to INSUFFICIENT_EVIDENCE. Mean response time uses the wall-clock duration logged by the experiment script. Because the answers are short, the automatic labels were also inspected against the output file. These measures are appropriate for this pilot, but they are not a replacement for claim-level human evaluation on a larger corpus.

5. RESULTS

	Correct	Hallucination	Abstention
	30/30	0/30	5/30
	29/30	0/30	6/30
	30/30	0/30	5/30
	29/30	0/30	6/30

Table -1: Recorded outcomes for 30 questions per configuration.

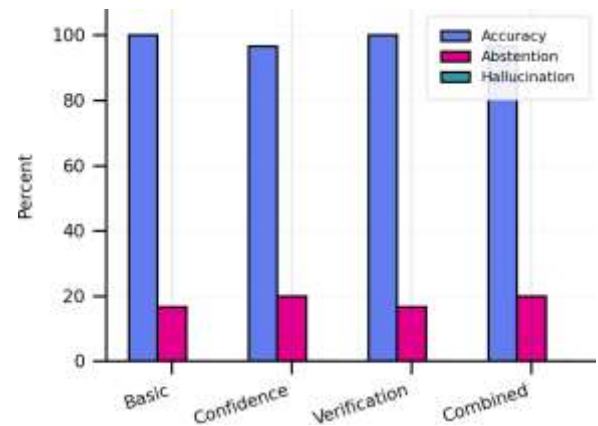


Fig -2: Accuracy, abstention, and observed hallucination rates.

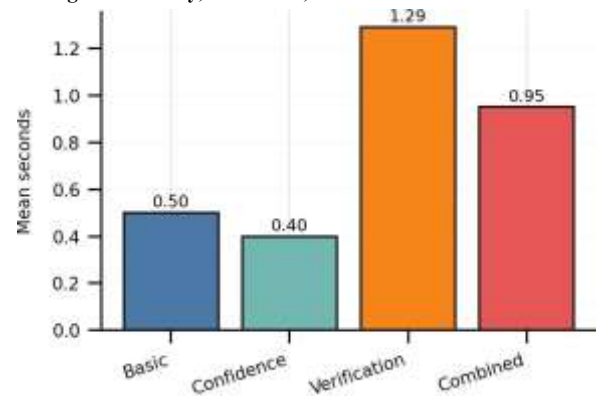


Fig -3: Mean response time measured during the run.

5.1 Correctness and refusals

Basic RAG and Verification-Based RAG produced a correct final outcome for every benchmark question. The five absent questions were answered with INSUFFICIENT_EVIDENCE, so the Basic configuration already behaved conservatively under the evidence-only prompt. Confidence-Aware and Combined RAG each lost one answerable case. For Q21, which asked about the purpose of FAISS, the relevant passage was ranked first, but the score was only 0.261. The fixed 0.55 threshold rejected the query even though the retrieved evidence contained the answer. This result is a false refusal, not a hallucination.

The absence of recorded hallucinations requires careful interpretation. The benchmark contains short passages and direct questions, and the prompt explicitly prohibits use of outside knowledge. The result therefore shows safe behaviour in this controlled setting; it does not establish a general zero-hallucination capability. A harder corpus with competing passages, long answers, or multi-step reasoning may produce a different ordering of methods.

5.2 Time cost

Confidence-Aware RAG was fastest on average because six low-score questions terminated before an LLM request. Verification-Based RAG was slowest because all 30 questions required both answer generation and verification. Combined RAG reduced some of that overhead by stopping six questions at the confidence gate. The timings show that a safety layer is an engineering choice with measurable cost rather than a free improvement.

6. INTERPRETATION

The two safeguards address different risks. A confidence gate asks whether retrieval appears strong enough to continue. A verifier asks whether the answer is actually licensed by the retrieved text. The first

can save calls and reduce unsupported answering; the second can detect excess claims after generation. In this experiment, the verifier maintained full benchmark coverage, while the gate introduced one avoidable refusal.

The Q21 case is the most useful outcome of the study. A raw embedding score is not a calibrated probability of correctness. Score distributions vary with wording, passage length, topic, and embedding model. A production system should tune the threshold on representative data, examine the margin between the first and second result, or route low-score cases to a secondary evidence check rather than reject them automatically.

For low-risk internal search, a strict evidence-only Basic RAG prompt may offer adequate performance with lower latency. In contexts where unsupported claims carry greater consequences, verification may be justified despite slower response. Confidence gating remains useful when refusal is acceptable, but threshold selection should be treated as an empirical calibration problem.

7. LIMITATIONS

The study uses a small author-constructed dataset, one embedding model, one vector index, and one hosted compound model. The questions are mostly single-fact lookups and do not contain contradictory documents or adversarial instructions. Accuracy relies on reference terms rather than independent human raters. Generation and verification use the same provider, which can create correlated judgments. The compound service may also route requests internally, limiting exact reproducibility. Consequently, the reported percentages describe this run only.

8. CONCLUSION

A lightweight RAG experiment compared a baseline against confidence gating, answer verification, and a combined pipeline. Basic and Verification-Based RAG returned 30 correct outcomes, while the confidence-gated methods returned 29 because one relevant passage fell below the selected threshold. No unsupported final answer was observed. Verification added latency but preserved coverage; confidence gating reduced calls but required calibration. The next stage should evaluate public QA benchmarks, conflicting evidence, longer answers, model-diverse verification, and human claim-level scoring before any enterprise deployment decision.

ACKNOWLEDGEMENT

The implementation used FAISS, Sentence Transformers, pandas, and the Groq API. Generative AI assistance was used during drafting and language revision. The authors remain responsible for checking citations, experimental records, originality, institutional affiliation, journal policy, and final submission content.

REFERENCES

- [1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.
- [2] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, Art. 248, 2023, doi: 10.1145/3571730.
- [3] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Transactions on Information Systems*, 2025, doi: 10.1145/3703155.
- [4] A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," *Proc. ICLR*, 2024.
- [5] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective Retrieval Augmented Generation," *arXiv:2401.15884*, 2024.

- [6] S. Dhuliawala et al., "Chain-of-Verification Reduces Hallucination in Large Language Models," *Findings of ACL*, pp. 3563-3578, 2024.
- [7] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," *arXiv:2309.15217*, 2023.
- [8] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proc. EMNLP*, pp. 6769-6781, 2020.
- [9] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535-547, 2021.
- [10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proc. EMNLP-IJCNLP*, pp. 3982-3992, 2019.
- [11] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection," *Proc. EMNLP*, pp. 9004-9017, 2023.
- [12] S. Min et al., "FactScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation," *Proc. EMNLP*, pp. 12076-12100, 2023.