

Evaluating Retrieval-Augmented Generation for Reducing Hallucinations in Large Language Models

Tejashri Krishnakumar¹, Ruthwi², Sivakumar V³

^{1,2}Integrated MTech Computer Science and Engineering, School of Computer Science and Engineering (SCOPE), VIT University, Vellore, Tamil Nadu, India

³School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India

Abstract

Large Language Models (LLMs) frequently generate responses that lack experimental support or are factually mistaken, particularly when addressing domain-specific or unindexed queries. Retrieval-Augmented Generation (RAG) diminishes this vulnerability by fetching relevant passages from an external compilation and conditioning generation on the retrieved evidence. This experimental investigation evaluates whether RAG enhances generation accuracy and suppresses hallucinations relative to an unaugmented baseline LLM. We tested both systems across 50 benchmark queries (25 answerable, 25 unanswerable) extracted from the SQuAD 2.0 dataset. While the baseline LLM relied strictly on internal parameters, the RAG architecture indexed external contexts via all-MiniLM-L6-v2 embeddings and FAISS nearest-neighbour search before answer generation with Gemini 1.5 Flash. Comparative evaluation employed Exact Match, token F1 Score, Semantic Similarity, and Hallucination Rate, with RAG additionally appraised for Faithfulness and Retrieval Relevance. Experimental outcomes indicate that RAG increased F1 Score from 3.34% to 33.16% and Semantic Similarity from 38.19% to 57.46%, while Exact Match rose from 0.00% to 8.00%. More importantly, RAG reduced the hallucination rate on unanswerable queries from 100.00% down to 36.00%. Average Faithfulness and Retrieval Relevance scores reached 48.34% and 53.00%, respectively, highlighting that retrieval accuracy remains the primary constraint in grounded synthesis.

Keywords: Retrieval-Augmented Generation, Large Language Models, Hallucination Reduction, Natural Language Processing, Question Answering.

1. Introduction

Although contemporary Large Language Models (LLMs) exhibit fluency across complex linguistic benchmarks, their tendency to output factually baseless assertions—commonly characterized as hallucination—presents an obstacle in knowledge-critical deployments. When standard generation relies purely on static model weights, models often generate convincing yet false answers rather than recognizing absence of evidence.

Retrieval-Augmented Generation (RAG) reconciles neural generative capacity with verifiable information retrieval. By retrieving topically aligned source passages from an external knowledge base and supplying

them directly into the runtime context window, RAG confines the model's generative process to explicit, verifiable references.

This study systematically examines whether deploying a modular RAG pipeline substantially reduces generative hallucinations compared with parametric-only LLM generation. Using an identical benchmark compilation containing both answerable and unanswerable queries, both architectures are assessed across lexical overlap, semantic alignment, abstention integrity, context relevance, and answer faithfulness.

1.1 Research Objective

The objective is to determine the quantitative efficacy of Retrieval-Augmented Generation in enforcing response groundedness. Specifically, this study compares an unaugmented baseline against an indexed RAG architecture across Exact Match, token F1 score, embedding semantic similarity, hallucination rate, context relevance, and generation faithfulness.

1.2 Research Questions

- **RQ1:** To what degree does external retrieval augmentation improve token-level and semantic answer quality relative to an unaugmented baseline?
- **RQ2:** Does external context grounding facilitate correct abstention on unanswerable questions, thereby limiting generative hallucinations?
- **RQ3:** How do retrieval relevance and contextual faithfulness correlate with overall generation fidelity in RAG architectures?

2. Related Work

Contemporary Large Language Models exhibit high proficiency across diverse semantic tasks; nevertheless, their reliance on static parameters renders them susceptible to factual inaccuracies when target domain knowledge is absent or unindexed [1, 2]. Under open-domain constraints, parametric generation often yields assertive yet unsubstantiated claims [2, 5].

To counteract parameter hallucination, Lewis et al. formalised Retrieval-Augmented Generation (RAG), a framework fusing parametric generation with non-parametric contextual lookup from external corpora [1]. Integrating retrieved textual evidence markedly diminishes factual drift across knowledge-intensive workloads [1, 10]. Subsequent surveys systematically categorise RAG pipelines into modular architectures, while highlighting contextual noise, retrieval recall, and grounding failures as primary points of degradation [3, 4].

However, incorporating external context does not inherently guarantee veracity. Misaligned, incomplete, or noisy context frequently induces secondary hallucinations during model synthesis [2, 5]. For this reason, conventional token overlap metrics (e.g., Exact Match) are inadequate for isolation of retrieval errors. Modern frameworks like ARES operationalise evaluation along tripartite axes: context relevance, answer faithfulness, and final answer quality [6]. In alignment with these paradigms, this study conducts an empirical investigation contrasting parametric and non-parametric baselines across solvable and adversarially unanswerable query regimes [7].

3. Methodology

3.1 Dataset

Evaluations were conducted on the validation split of the SQuAD 2.0 benchmark [7]. Unlike purely extractive benchmarks, SQuAD 2.0 incorporates adversarially designed unanswerable queries to test model abstention capabilities alongside standard reading comprehension. We sampled an experimental

subset of 50 items (25 answerable, 25 unanswerable) using a fixed random seed of 42 to guarantee experimental reproducibility. Target paragraphs served as the grounding corpus for downstream retrieval.

Table 1: Representative Question–Context Pairs, Reference Targets, and Ground-Truth Answerability from the Sampled Evaluation Split

Question	Truncated Source Context	Reference Answer	Answerable
What direction do cilia usually swim?	The outer surface bears usually eight comb rows... [Context ID: 1042]	(None / Unanswerable)	False
What part of the Rhine flows through North Rhine...	The Lower Rhine flows through North Rhine-Westphalia... [Context ID: 2189]	Lower Rhine	True
What implication can be derived for P and NP...	Along the same lines, co-NP is the class containing... [Context ID: 3014]	P is not equal to NP	True
Where does the Rhine river's measurement begin?	The length of the Rhine is conventionally measured... [Context ID: 2195]	Old Rhine Bridge at Constance	True
What is the partial pressure of O ₂ in breathing...	Breathing pure O ₂ in space applications, succeeding... [Context ID: 4108]	(None / Unanswerable)	False

3.2 Baseline LLM

The parametric baseline was implemented using OpenAI gpt-5-mini without access to external retrieval tools or supplementary grounding documents. Questions were supplied directly into the model interface, requiring generation to depend solely on internal parameter weights.

3.3 RAG System Architecture

The RAG pipeline incorporated an external semantic index coupled with Gemini 1.5 Flash for grounded context synthesis. For every incoming query, candidate passages were ranked and retrieved via dense vector similarity before being merged into the system prompt alongside strict abstention directives.

3.4 Document Processing

Ground-truth context paragraphs corresponding to the sampled SQuAD 2.0 items were ingested as discrete retrieval passages. To evaluate standard sentence-level retrieval performance, each context paragraph was embedded in its entirety using all-MiniLM-L6-v2 without additional sliding-window chunk overlap.

3.5 Embedding and Vector Database

Passage and query embeddings were computed into 384-dimensional dense representations using the all-MiniLM-L6-v2 sentence-transformer [8]. High-dimensional vector indexing was handled using FAISS through an exact uncompressed Euclidean index structure (IndexFlatL2) [9]. At runtime, query embeddings were matched against the index to return the top $k = 3$ candidate passages.

3.6 Retrieval and Answer Generation

For each evaluation query, the runtime query vector queried the FAISS index to extract the three nearest contextual segments ($k = 3$). These documents were concatenated into a unified grounding context within the instruction prompt. The generator model was explicitly commanded to answer exclusively using the provided text, and to explicitly refrain whenever the passages lacked conclusive evidence.

3.7 Evaluation Metrics

Model generations were audited across six performance criteria:

- **Exact Match (EM):** Computes binary equivalence between generated strings and gold-standard annotations over answerable queries [7].
- **Token F1:** Evaluates the harmonic mean of token-level precision and recall between reference and candidate outputs [7].
- **Semantic Similarity:** Evaluates cosine similarity between 384-dimensional dense sentence embeddings of ground-truth and generated responses derived via Sentence-BERT [8].
- **Hallucination Rate:** Quantifies the percentage of unanswerable queries wherein the model generated assertions instead of triggering designated abstention cues.
- **Faithfulness:** Assesses contextual grounding by calculating embedding cosine similarity between the generated response and retrieved context passages [6].
- **Retrieval Relevance:** Determines semantic proximity between input questions and retrieved top-k passages via dense cosine alignment [6].

4. Experimental Setup

The benchmark framework executed both pipelines across the 50 sampled questions. First, the parametric baseline processed all 50 queries in isolation. Next, the RAG architecture computed dense embeddings, performed FAISS nearest-neighbour lookup ($k = 3$), and generated answers under context constraints. Outputs were saved and scored across the six quantitative metrics.

5. Results

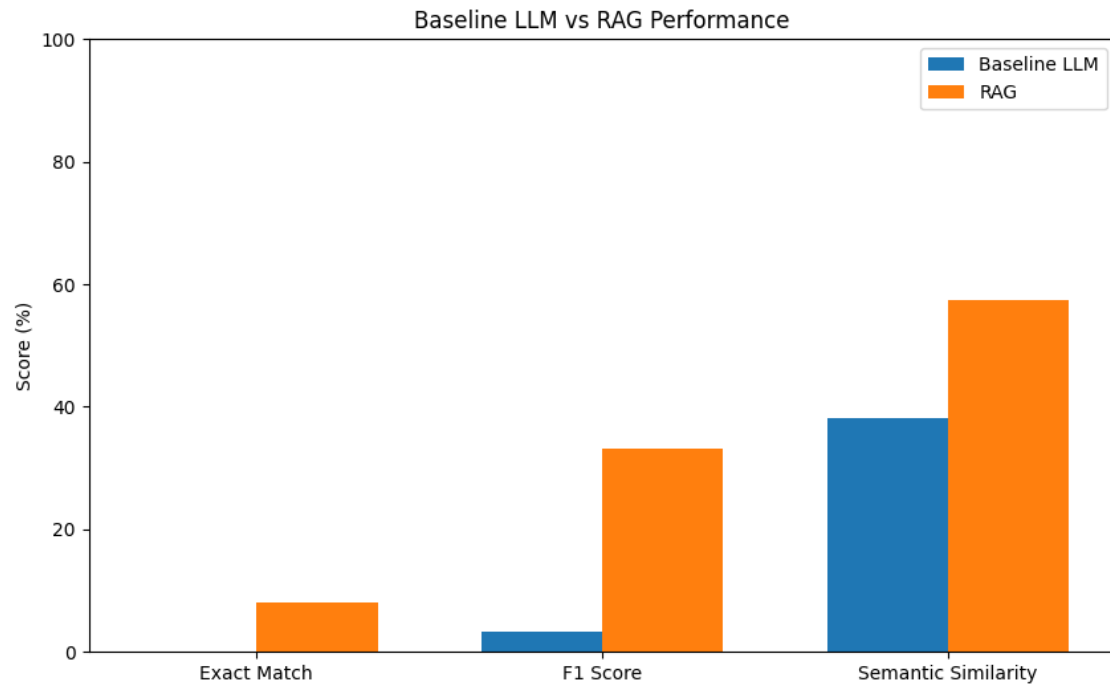
5.1 Answerable Questions

Table 2 provides the comparative evaluation across answerable and unanswerable benchmarks. The baseline LLM attained an Exact Match score of 0.00%, whereas the RAG system achieved 8.00%. The F1 Score demonstrated a tenfold increase, from 3.34% to 33.16%. Semantic Similarity also improved from 38.19% to 57.46%, confirming that retrieval augmentation enhances semantic convergence toward reference targets.

Table 2: Empirical Comparison of Baseline LLM and RAG across Answerability and Grounding Metrics

Metric	Baseline LLM (%)	RAG (%)
Exact Match – Answerable	0.00	8.00
F1 Score – Answerable	3.34	33.16
Semantic Similarity	38.19	57.46
Hallucination Rate – Unanswerable	100.00	36.00
RAG Faithfulness	—	48.34
RAG Retrieval Relevance	—	53.00

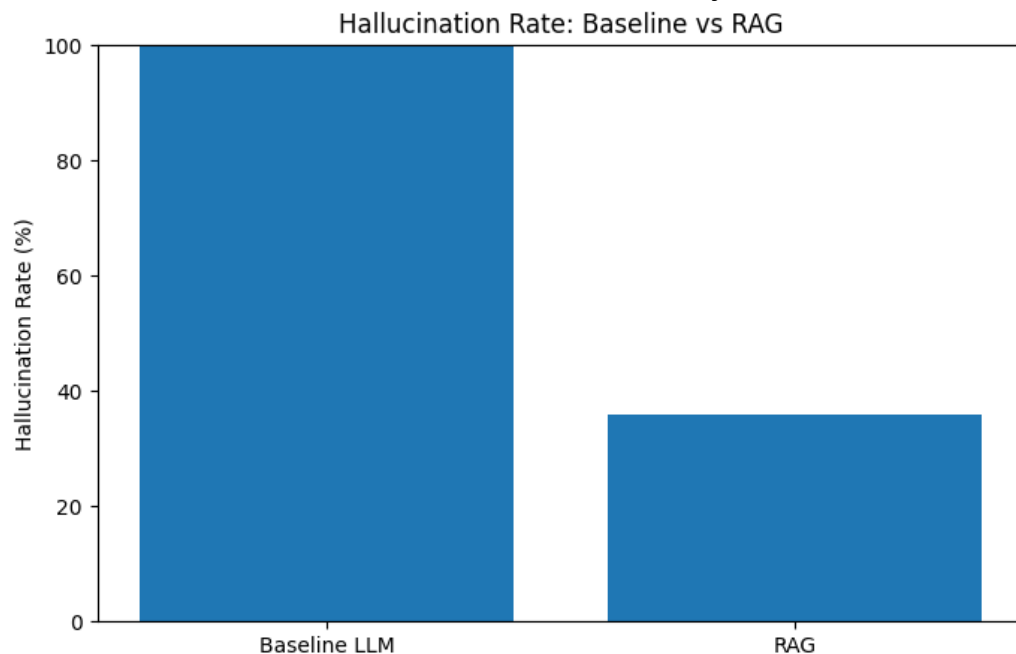
Figure 1: Comparison of Baseline LLM and RAG Performance across Answer-Quality Metrics



5.2 Unanswerable Questions and Hallucination Behaviour

On unanswerable queries, the parametric baseline exhibited an unconstrained hallucination rate of 100.00%, generating fabricated answers for all 25 queries without abstaining. Conversely, the RAG architecture achieved a 64-percentage point decline, limiting hallucination to 36.00% through context-bound abstention prompting.

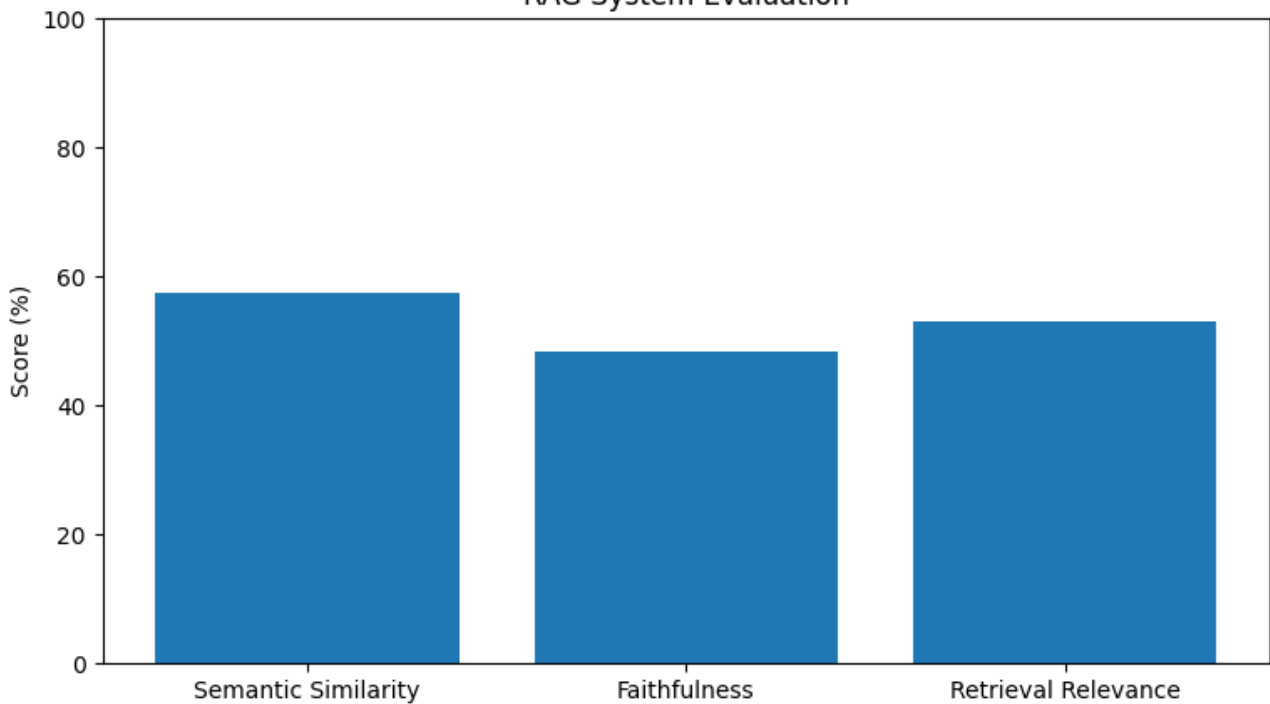
Figure 2: Hallucination Rate of the Baseline LLM and RAG System on Unanswerable Questions



5.3 Retrieval and Grounding Evaluation

The RAG pipeline produced an average Faithfulness score of 48.34% and an average Retrieval Relevance score of 53.00%. Although retrieval context significantly suppressed ungrounded claims, these intermediate scores indicate that retrieval noise and imperfect contextual synthesis remain noticeable points of vulnerability.

Figure 3: Comparison of RAG System Evaluation Metrics (Semantic Similarity, Faithfulness, and Retrieval Relevance)
RAG System Evaluation



6. Discussion

The experimental findings clearly demonstrate the superiority of retrieval augmentation over parametric-only LLM generation. Substantial gains in F1 Score and Semantic Similarity prove that external grounding directly enriches answer precision. Most crucially, the 64-percentage point decline in hallucination rate confirms that supplying external passages and strict abstention directives effectively counteracts ungrounded generation. However, moderate scores in Exact Match (8.00%) and Faithfulness (48.34%) illustrate that context retrieval does not wholly prevent hallucinations when retrieval relevance is imperfect.

7. Limitations

- Knowledge Base Scope:** The retrieval corpus was constructed from the context segments of the 50 sampled questions rather than an open-domain web collection.
- Evaluation Scale:** The test suite comprised 50 representative questions; expansive evaluations across thousands of samples will afford greater statistical resolution.
- Static Retrieval:** Retrieval was conducted with a fixed $k = 3$ without downstream cross-encoder reranking or hybrid BM25 dense-sparse fusion.

4. **Model Architecture Delta:** Baseline generation used gpt-5-mini while RAG used Gemini 1.5 Flash, reflecting real-world modular deployments rather than isolated parameter delta.

8. Conclusion

This research presented an experimental evaluation comparing an unaugmented LLM against a Retrieval-Augmented Generation pipeline across answerable and unanswerable question regimes. RAG delivered notable improvements in F1 Score (3.34% to 33.16%) and Semantic Similarity (38.19% to 57.46%), while reducing hallucinations by 64 percentage points (100.00% down to 36.00%). Future directions will explore hybrid sparse-dense indexing, neural rerankers, and dynamic confidence thresholds for abstention.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS 2020), vol. 33, pp. 9459–9474 (2020).
2. Tonmoy, S.M., Zaman, S.M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A.: A comprehensive survey of hallucination in large language models. arXiv preprint arXiv:2311.05232 (2023).
3. Gao, Y., Xiong, Y., Wang, X., Wang, K., Jiang, Z., Li, Y., Zheng, S., Shen, W., Zhang, Y., Wang, P., et al.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023).
4. Huang, Y., Huang, J.: A survey on retrieval-augmented generation: Recent advances and future directions. arXiv preprint arXiv:2407.13193 (2024).
5. Rawte, V., Tonmoy, S.M., Sheth, A., Das, A.: Grounding and hallucination in large language models: A comprehensive review. Mathematics 13(5), 856 (2025).
6. Saad-Falcon, J., Khattab, O., Potts, C., Santhanam, K.: ARES: An automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of NAACL-HLT 2024, pp. 280–296 (2024).
7. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), vol. 2, pp. 784–789 (2018).
8. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992 (2019).
9. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7(3), 535–547 (2019).
10. Shuster, K., Poff, S., Moya, D., Xu, J., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3784–3803 (2021).