Adaptive RAG Framework for Reducing LLM Hallucinations

Yash Suhas Shete
Shubhangi Pratik Bomble
Department of Master of Computer Application
JCEI's Jaihind Institute of Management and Research
Vadgaon Sahani, India
Email: sheteyash001@gmail.com, gadeshubhangi9@gmail.com

Abstract: **Large Language Models (LLMs) show strong ability in natural language processing, text generation, and question answering. They frequently produce hallucinations, however: responses that are inaccurate, misleading, or not backed by evidence. Such errors weaken the reliability of LLM-based applications, especially in fields where factual accuracy is essential. Retrieval-Augmented Generation (RAG) tackles this problem by bringing external knowledge sources into the response-generation process. Even so, conventional RAG systems can still hallucinate because they retrieve irrelevant documents, receive insufficient context, or fail to integrate the retrieved knowledge effectively. This research proposes an Adaptive Retrieval-Augmented Generation (Adaptive RAG) framework that reduces hallucinations by adjusting retrieval and generation to the complexity of each user query and the relevance of the retrieved information. The framework comprises query analysis, adaptive retrieval strategies, document relevance evaluation, context refinement, and response verification. Depending on the needs of the query, it decides whether further retrieval is required, filters out irrelevant material, and checks the generated response against the retrieved evidence before presenting the final answer. The framework is to be assessed with factual accuracy, hallucination rate, retrieval precision, answer relevance, and response latency, through comparative experiments against a standalone LLM and a standard RAG system. The expected outcome is greater factual reliability and more evidence-based responses without a loss of computational efficiency. Overall, this research aims to offer a practical route to more trustworthy LLM-powered applications.**

Keywords: **Large Language Models; Retrieval-Augmented Generation; Adaptive RAG; Hallucination Reduction; Information Retrieval; Factual Accuracy; Response Verification.**

I. INTRODUCTION

Artificial Intelligence (AI) has advanced rapidly in recent years, most visibly through Large Language Models (LLMs). These models can interpret natural language, produce human-like text, answer questions, summarize documents, and power a wide range of intelligent applications. LLMs are now used in education, healthcare, software development, research, and customer support. Despite these strengths, they suffer from a serious problem called hallucination, in which the model produces content that is factually wrong, misleading, or lacking reliable support [5]. Hallucinations arise from gaps in training data, incomplete knowledge, ambiguous user queries, and the probabilistic nature of text generation. These weaknesses undermine the trustworthiness of AI-based systems.

Retrieval-Augmented Generation (RAG) aims to improve the factual accuracy of LLMs by combining external knowledge retrieval with text generation [1], [7]. Rather than depending only on what was learned during training, a RAG system fetches relevant documents from external sources and uses them as context when producing an answer, which improves the relevance and factual grounding of responses. Traditional RAG systems nevertheless have drawbacks: they may retrieve irrelevant documents, miss important information, rely on outdated knowledge, and generate answers that the retrieved context does not fully support.

This research proposes an Adaptive Retrieval-Augmented Generation (Adaptive RAG) framework for reducing LLM hallucinations. The framework makes retrieval and generation more flexible by tailoring its operations to each query. It consists of query analysis, adaptive retrieval, document relevance assessment, context refinement, response generation, and factual verification.

First, the system analyzes the user's query to estimate its complexity and information needs, and on that basis chooses a suitable retrieval strategy. The retrieved documents are then assessed for relevance and usefulness, and irrelevant or redundant content is removed. The remaining context is passed to the LLM, which generates an evidence-based response. A verification component then checks whether the retrieved information supports the generated answer. If the evidence is insufficient or the response contains unsupported claims, the framework can start another retrieval cycle or state its uncertainty. This iterative process is intended to raise factual reliability while avoiding unnecessary retrieval.

A. Problem Statement

Traditional RAG systems improve LLM performance, but they do not eliminate hallucinations. Most conventional implementations follow a largely fixed retrieval process, regardless of how complex the query is or how good the retrieved information turns out to be. Simple queries may trigger unnecessary retrieval, while complex queries may call for extra searches, better document ranking, or several retrieval iterations.

In addition, an LLM may misread the retrieved information or add unsupported claims while composing its response. An intelligent, adaptive framework is therefore needed, one that controls retrieval dynamically, assesses the quality of the retrieved documents, and verifies generated responses against the available evidence.

B. Objectives

The main objective of this research is to design an adaptive RAG framework that reduces hallucinations in LLM-generated responses. The specific objectives are:

- To examine the causes and limitations of hallucinations in Large Language Models.
- To study how conventional Retrieval-Augmented Generation systems work and where they fall short.

- To develop an adaptive retrieval mechanism driven by query complexity and document relevance.
- To raise the quality of retrieved information through document filtering and context refinement.
- To include response verification techniques that detect unsupported or potentially inaccurate statements.
- To evaluate the proposed framework in terms of factual accuracy, hallucination rate, retrieval precision, answer relevance, and response latency.
- To compare the proposed framework with a standalone LLM and a conventional RAG system.

C. Contributions

- Proposes an adaptive RAG framework that adjusts retrieval to query complexity.
- Combines document relevance evaluation and context refinement to remove irrelevant or redundant evidence.
- Adds a verification step that checks responses against retrieved evidence and triggers re-retrieval or an uncertainty statement.
- Defines an evaluation plan that compares the framework with a standalone LLM and a conventional RAG system.

II. LITERATURE REVIEW

Lewis et al. [1] introduced Retrieval-Augmented Generation (RAG), which pairs a pre-trained language model with an external document retrieval component. This design lets the model draw on information fetched from a knowledge source while it generates a response, easing the limits of depending solely on knowledge stored in model parameters, especially for knowledge-intensive tasks such as question answering. The work laid an important foundation for combining information retrieval with natural language generation. Its effectiveness, however, depends on the quality of the retrieved documents and on the model's ability to use them correctly. Irrelevant or incomplete context can still produce inaccurate or unsupported responses.

Asai et al. [2] proposed Self-RAG, in which the language model learns to retrieve passages on demand and to reflect on both the retrieved passages and its own output using special reflection tokens. This improves factuality and citation quality compared with standard retrieval-augmented models. Jeong et al. [3] addressed the rigidity of fixed pipelines with Adaptive-RAG, which uses a classifier to estimate query complexity and then routes each query to no retrieval, single-step retrieval, or multi-step retrieval. Their results show that matching the strategy to the query balances accuracy against efficiency.

Evaluation and hallucination detection form a complementary line of work. Es et al. [4] introduced RAGAS, a set of reference-free metrics for assessing a RAG pipeline, including how faithful the answer is to the retrieved context and how relevant the answer and context are to the question. Ji et al. [5] surveyed hallucination in natural language generation and distinguished content that contradicts the source (intrinsic) from content that cannot be verified from it (extrinsic). Manakul et al. [6] proposed SelfCheckGPT, which flags likely hallucinations by sampling several responses and measuring their consistency, without needing external resources. Gao et al. [7] reviewed the field of RAG and grouped its development into naive, advanced, and modular paradigms.

The literature shows that conventional RAG, adaptive retrieval, self-reflective generation, and automated evaluation each help make LLM-based systems more reliable, yet important challenges remain. A retrieval system may select documents that relate to a query but do not contain enough evidence to answer it. A language model may also make unsupported statements even when relevant documents are available. Moreover, repeated retrieval raises computational cost and response latency unless it is properly controlled.

This research therefore proposes an Adaptive RAG framework that brings together query complexity analysis, document relevance assessment, context refinement, evidence sufficiency checking, and response verification in a single workflow. When the evidence is insufficient, the system can retrieve again or express uncertainty instead of producing an unsupported answer. The aim is to test whether combining these mechanisms improves factual reliability while keeping computational cost acceptable.

TABLE 1 COMPARATIVE ANALYSIS OF EXISTING STUDIES

Research Study	Main Contribution	Identified Limitation
Lewis et al. (2020)	Combines external retrieval with language generation.	Answer quality depends on the retrieved evidence and how it is interpreted.
Asai et al. (2024)	Combines adaptive retrieval with self-reflection to improve factuality.	Self-reflection alone does not guarantee factual correctness.
Jeong et al. (2024)	Chooses a retrieval strategy based on query complexity.	Query complexity alone does not establish the reliability of the evidence.
Es et al. (2023)	Provides metrics for evaluating RAG pipelines.	Evaluation scores alone cannot guarantee factual correctness.

III. METHODOLOGY

A. System Architecture

The proposed Adaptive RAG framework seeks to reduce hallucinations in LLMs by dynamically adjusting the retrieval, context processing, and response generation stages. Whereas conventional RAG systems generally follow a fixed retrieval pipeline, this framework selects its retrieval strategy according to query complexity, document relevance, and whether the available evidence is sufficient. It has six main components: query analysis, adaptive retrieval, document relevance evaluation, context refinement, response generation, and factual verification. Together they produce responses grounded in reliable retrieved information.

The architecture follows a sequential, iterative workflow. The user first submits a query, and the query analysis module examines its complexity and information needs. Using this analysis, the adaptive retrieval module decides whether retrieval is needed and picks a suitable strategy. The retrieved documents are then ranked by their relevance to the query. Next, the context refinement module removes irrelevant or redundant passages and prepares the most useful information for the LLM, and the response generation module produces an answer from this refined context.

Finally, the factual verification module checks whether the retrieved evidence supports the generated response. If the evidence is insufficient or the response contains unsupported claims, the system may repeat retrieval or return a response that clearly states the limits of the available information. In sequence, the stages are: user query input, query complexity analysis, adaptive retrieval strategy selection, document retrieval and ranking, context refinement, LLM-based response generation, evidence-based response verification, and finally either the answer or additional retrieval.

B. Adaptive Retrieval Mechanism

The adaptive retrieval mechanism decides how much external information a query needs. Simple queries that the language model can answer reliably on its own may skip retrieval, knowledge-intensive queries use a single retrieval step, and complex queries that need several pieces of evidence may trigger iterative retrieval.

The strategy is chosen using factors such as query complexity, the availability of relevant information, retrieval confidence, and evidence sufficiency. A query classifier can sort queries into simple, moderate, and complex groups, and each group is mapped to a suitable retrieval strategy. For example, a simple factual question may need only one document, whereas a multi-step analytical question may need several documents and multiple retrieval iterations. If the retrieved evidence does not adequately address the query, the system can reformulate the search query and retrieve more documents. This adaptive process aims to balance response accuracy against computational efficiency. Table 2 summarizes the components of the framework.

TABLE 2 METHODOLOGY OF THE PROPOSED ADAPTIVE RAG FRAMEWORK

Sr. No.	Component	Description	Expected Outcome
1	Query Input	Accepts the user's question or information request.	Provides input for processing.
2	Query Complexity Analysis	Classifies queries as simple, moderate, or complex.	Identifies the appropriate retrieval strategy.
3	Adaptive Retrieval	Selects no retrieval, single-step retrieval, or iterative retrieval based on query needs.	Improves retrieval efficiency.
4	Document Retrieval	Retrieves relevant documents from an external knowledge base using semantic search.	Supplies supporting information.
5	Context Refinement	Removes irrelevant or redundant passages and keeps the most useful content.	Provides focused, high-quality context.
6	Response Generation	Generates an answer from the refined context.	Produces an evidence-based response.
7	Factual Verification	Checks the answer against the retrieved evidence and triggers re-retrieval or an uncertainty statement when support is lacking.	Reduces unsupported claims.

C. Evaluation Plan

The framework will be evaluated in terms of factual accuracy, hallucination rate, retrieval precision, answer relevance, and response latency. Comparative experiments will be run against a standalone LLM and a conventional RAG system, so that any gain in factual reliability can be weighed against the computational cost of adaptive retrieval and verification.

IV. DISCUSSION

The proposed Adaptive RAG framework aims to improve the factual reliability of LLMs by dynamically adjusting retrieval and response generation. Conventional RAG systems retrieve external information to support answer generation, but their effectiveness depends on the relevance, completeness, and reliability of the retrieved documents. The proposed framework addresses these challenges by combining query complexity analysis, adaptive retrieval, document relevance evaluation, context refinement, and response verification in one workflow.

The framework selects its retrieval strategy to suit each query. Simple queries may need little retrieval, whereas complex questions may need several retrieval operations to gather enough evidence. This approach is expected to improve retrieval efficiency and lower the chance of answers built on incomplete or irrelevant context.

V. CONCLUSION AND FUTURE WORK

Large Language Models have shown considerable ability in natural language understanding and text generation. Their tendency to hallucinate, that is, to produce inaccurate, misleading, or unsupported information, nevertheless remains a major obstacle to reliability. RAG eases this problem by bringing external knowledge sources into response generation, but conventional systems can still hallucinate because of irrelevant retrieval, insufficient evidence, outdated information, and misinterpretation of the retrieved context.

This research proposes an Adaptive RAG framework that combines query complexity analysis, adaptive retrieval strategy selection, document relevance evaluation, context refinement, response generation, and evidence-based response verification. By adjusting retrieval to the needs of each query, the approach aims to improve the quality of retrieved information and avoid unnecessary retrieval. In addition, the verification mechanism is designed to detect unsupported claims and, when evidence is lacking, trigger further retrieval or communicate uncertainty. Several directions could extend this work:

- **Advanced query classification:** Machine-learning-based classifiers could select retrieval strategies more accurately, helping the system tell apart simple factual questions, complex multi-step queries, and questions that need additional external evidence.
- **Hybrid retrieval and source quality:** Combining keyword-based search with semantic vector search could improve retrieval across query types, while document credibility assessment, source verification, and timestamp-based filtering may reduce the influence of unreliable or outdated information.
- **Claim-level verification:** Individual statements in a response could be compared with the retrieved evidence, supported by automated citation generation and better uncertainty estimation to show when the evidence is too weak for a reliable answer.
- **Computational efficiency:** Adaptive stopping criteria, efficient document reranking, and controlled retrieval iterations could cut response latency and resource consumption with little effect on answer quality.
- **Broader evaluation:** Experiments across multiple datasets, language models, and application domains, together with ablation studies and human assessment, could reveal which components contribute most to hallucination reduction and how well the framework generalizes.

REFERENCES

[1]. P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[2]. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in International Conference on Learning

Representations (ICLR), 2024. [Online]. Available: https://openreview.net/forum?id=hSyW5go0v8

[3]. S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," in Proc. 2024 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2024, pp. 7036–7050, doi: 10.18653/v1/2024.naacl-long.389.

[4]. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv:2309.15217, 2023. [Online]. Available: https://arxiv.org/abs/2309.15217

[5]. Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Art. 248, pp. 1–38, 2023, doi: 10.1145/3571730.

[6]. P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 9004–9017. [Online]. Available: https://aclanthology.org/2023.emnlp-main.557/

[7]. Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[8]. S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective Retrieval Augmented Generation," arXiv:2401.15884, 2024. [Online]. Available: https://arxiv.org/abs/2401.15884

[9]. V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.

[10]. Z. Jiang *et al.*, "Active Retrieval Augmented Generation," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, doi: 10.18653/v1/2023.emnlp-main.495.

[11]. H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions," in Proc. 61st Annu. Meeting of the Association for Computational Linguistics (ACL), 2023, doi: 10.18653/v1/2023.acl-long.557.

[12]. S. Dhuliawala *et al.*, "Chain-of-Verification Reduces Hallucination in Large Language Models," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 3563–3578, doi: 10.18653/v1/2024.findings-acl.212.

[13]. N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," Trans. Assoc. Comput. Linguistics (TACL), 2024, doi: 10.1162/tacl_a_00638.

[14]. S. Min *et al.*, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[15]. J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 16, 2024, pp. 17754–17762, doi: 10.1609/aaai.v38i16.29728.