SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

A SYSTEMATIC REVIEW OF OPTIMIZED RETRIEVAL-AUGMENTED GENERATION FRAMEWORKS FOR REDUCING HALLUCINATIONS IN LARGE LANGUAGE MODELS

Jaipal Kumar[1], Dr. Shashank Swami[2]
[1]*MTech Student, Vikrant University Gwalior, MP.*
[2]*Assistant Professor, Department of Computer Science & Engineering, Vikrant University, Gwalior MP*

Article DOI: https://doi.org/10.36713/epra26155
DOI No: 10.36713/epra26155

ABSTRACT
Large Language Models (LLMs) have proven to be making great progress on natural language processing, but they tend to produce hallucinated or even inaccurate information. The systematic review is focused on optimized Retrieval-Augmented Generation (RAG) frameworks and how they enhance the reliability and factual truth of the outputs of LLM. The paper examines the architecture of the RAG systems, retrieval methods, and optimization methods like embedding tuning, query expansion, and reranking. It also surveys methods of evaluation by such metrics as factual consistency and faithfulness. The results show that the incorporation of outside knowledge recall massively decreases hallucinations and improves the grounding in context. Although there are difficulties that are associated with scalability and knowledge management, optimized RAG frameworks present a good pathway towards creating trustworthy and evidence-based generative AI systems.
KEYWORDS: *Large Language Models; Retrieval-Augmented Generation; Hallucination Reduction; AI Reliability.*

1.INTRODUCTION

Large Language Models (LLM) is one of the most revolutionary technologies in the field of artificial intelligence, which allows to develop advanced natural language comprehension and produce texts in various fields, including healthcare, education, legal services, finance, customer support, and knowledge management. The trained on large-scale textual corpora and usually built on transformer-based architectures, modern LLMs exhibit compelling reasoning and content generation, summarization, and conversational interaction capabilities. Their high rate of adoption in industries is because of the ability to scale, flexibility and automating of knowledge intensive tasks. Nevertheless, even with the impressive performance, LLMs are probabilistic systems which produce answers based on the learned patterns instead of the proven facts, which is why it is problematic because such a system is highly vulnerable to a well-known issue called hallucination.

Hallucinations of LLMs are the creation of information that is either factually false, created or unsubstantiated. These mistakes can take the form of misplaced references, fabricated information or inconsistent results. This may seriously challenge the user trust and reliability especially in critical areas like in healthcare, legal judgment and academic studies. The fact that hallucinated content has an adverse impact on downstream usage of information that is accurate and verifiable also shows the need to create mechanisms that improve the factual consistency and reliability of AI-generated responses.

In response to such issues, factual grounding and explainability in generative AI systems is in growing demand. Not only should reliable systems put forward coherent results, but they should also be able to offer evidence-based results that can be checked with trusted sources of knowledge. In that regard, Retrieval-Augmented Generation (RAG) has proven to be a feasible way of reducing hallucinations through the combination of external knowledge retrieval processes and generative models. RAG systems do not necessarily use solely the knowledge that is static during the pre-training stage; they can access the appropriate document or data in external repositories and use them in the generation process, thus enhancing the accuracy of facts and transparency.

RAG frameworks are also a major change of the design of language models, no longer static pre-trained knowledge but dynamic and context-aware information retrieval. The initial language models only used stored parameters and current optimized RAG models use more sophisticated retrieval mechanisms which include dense vector search, hybrid retrieval, reranking, and adaptive retrieval systems. Such innovations enable the LLMs to retrieve current and domain-specific information in the inference process, making them less likely to provide outputs that have been hallucinated and more context-relevant.

Since there is increasing research interest in the optimization of RAG systems to hallucinations reduction, systematic review of hitherto available methodologies, performance measurement criteria and architectural advancement is crucial. The main aim of the systematic

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

review is to examine the recent advancements in the sphere of optimized Retrieval-Augmented Generation systems, detect the successful methods of the hallucination reduction, and evaluate their performance in various application areas. The review will also capture the existing problems, gaps in the research and future of developing more reliable and trustworthy systems that are based on LLM.

Table 1: Summary of Recent Research on Hallucination Reduction and RAG Optimization

Author(s) & Year	Study Focus	Methodology	Key Findings	Contribution
Abdelghafour et al. (2024) [1]	Hallucination mitigation techniques in LLMs	Review study	Knowledge grounding and prompt strategies reduce hallucinations	Classified major mitigation approaches
Ahadian & Guan (2025) [2]	Survey of hallucination in foundation models	Literature survey	Data bias and incomplete knowledge are major causes	Provided taxonomy of hallucination types
Alansari & Luqman (2025) [3]	Comprehensive hallucination analysis	Systematic review	Factual and logical hallucinations are widespread	Identified detection and evaluation methods
Amugongo et al. (2025) [4]	RAG in healthcare applications	Systematic review	RAG improves factual reliability	Validated domain-specific use of RAG
Andriopoulos & Pouwelse (2023) [5]	Knowledge augmentation for LLMs	Survey	External knowledge reduces hallucination	Linked retrieval integration with trustworthiness

2. FUNDAMENTALS OF LARGE LANGUAGE MODELS AND HALLUCINATIONS

Large Language Models (LLM) are a category of natural language processing systems based on deep learning that generate and comprehend human text through the statistical association of large corpora [6]. Their ubiquitous use in fields like education, healthcare, legal analytics, and enterprise automation has necessitated the need to learn about the way they work, as well as their corresponding limitations. One of the major issues related to such models is the hallucination phenomenon whereby outputs produced by them might seem grammatically correct yet unrelated to reality or logical support. The section includes the basic architecture of LLMs, the nature and the classification of hallucinations, and the main causes that lead to such errors.

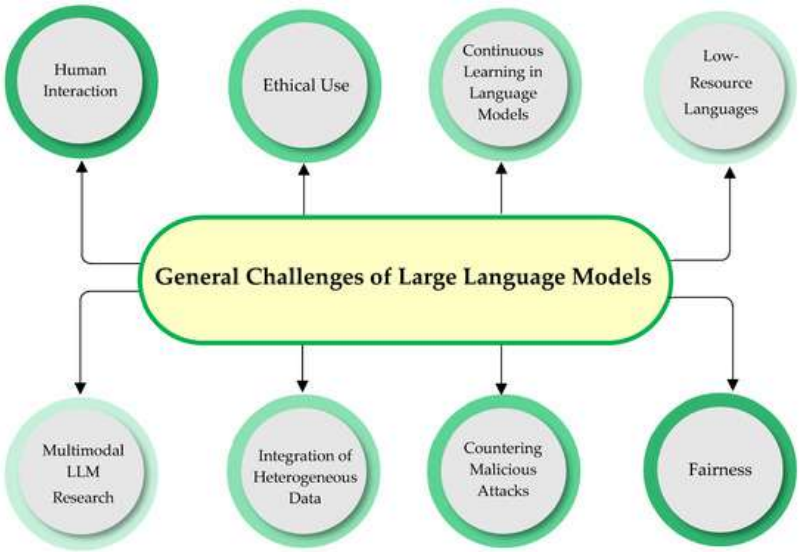


Figure 1: Comprehensive Taxonomy of Large Language Models and Their Associated Challenges and Mitigation Strategies [7]

2.1 Architecture and Training of LLMs

Transformer Architecture: In the majority of the modern LLMs, the transformer architecture is the one that uses self-attention to model contextual relationships between tokens in a sequence [8]. Compared to previous recurrent or convolutional designs, transformers compute input data simultaneously, which supports the ability to train on large datasets. The self-attention mechanism enables the model to balance the significance of various words against each other, hence, obtaining long-range connections and semantic patterns. The multi-head attention, position encoding, and multi-head encoder-decoding block feature together is used to enable advanced language comprehension and production.

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

Pre-training and Fine-Tuning Paradigm: LLMs are generally trained in two steps, including pre-training and fine-tuning. In pre-training, the model is trained to predict masked or subsequent tokens on large scale unlabeled datasets, especially in large unlabeled datasets. This step allows acquisition of syntactic structures and semantic relationships and context. Then domain or task-specific datasets are fine-tuned to specialize the model to specialized tasks like question-answering, summarization or dialogue-generation. Although this is flexible and scalable, the fact that probabilistic pattern learning is used instead of explicit fact verification is one of the reasons that may lead to the inaccuracy.

Limitations of Knowledge Representation: Although any LLM has a large parameter space and vast training data, it does not explicitly and structurally represent knowledge and can not be verified. Rather, the implicit representation of knowledge is in distributed vectors in model weights. The implicit storage is associated with a number of limitations such as the inability to dynamically retrieve new information, the outdated information, and the vulnerability to creation of plausible and false information. Further, lack of direct grounding on outside bodies of knowledge may lead to discrepancies in answering area particular or dynamic queries.

2.2 Nature and Types of Hallucinations

The hallucinations of the LLM outputs are syntactically fluent but lack factual support or are logically inconsistent. The hallucinations may take various forms based on the context and information that is available at the time of the generation.

Intrinsic vs Extrinsic Hallucinations: Intrinsic hallucinations happen when the content created conflicts with information or distorts information that exists in the input or context provided. An example of such a scenario is a summarization model that can add information that was not present in the original text. In contrast, extrinsic hallucinations are associated with the production of completely imaginary or unprovable information that is not supported by any external or contextual data. The two forms compromise reliability, especially where a knowledge intensive application is required.

Invented Facts, Misplaced Citation, Logical Fallacies: Invention of references that are not there, wrongful claim of authorship, made up numbers, and imperfect chains of reasoning are all common hallucinatory phenomena. These outputs might be credible because of the fluent use of language but they can be misleading to the users particularly in areas where the decision is actually critical. There might also be logical hallucinations when the model will come up with conclusions that are not to be readily determined as a result of the evidence presented and this will decrease the credibility and the interpretability of the results provided by AI.

2.3 Causes of Hallucination

Multiple interplaying factors involving data, model structure and inference methods affect the occurrence of hallucinations.
Data Bias: Training data are also usually biased, incomplete or even noisy. As LLMs are trained on these datasets to learn statistical associations, they can reproduce errors or overgeneralization of patterns that are not always universally true [9]. This may result in false or biased answers.

Incomplete Knowledge: Since the text models are based on the data they have stored in the course of pre-learning, they can be unaware of recent events or facts specific to the domain that they have not seen in the training set. In situations where there are gaps of knowledge, the model might only produce the plausible answers but incorrect ones as opposed to stating the uncertainty.

Timeliness: poor prompts or ambiguity can lead to failure to gain the right meaning of user intent. The lack of clear contextual guidance can make the model make use of generalized language patterns that make them more likely to produce hallucinated or irrelevant outputs.
Decoding Strategies: Generation parameters: Temperature, top-k sampling and nucleus sampling affect output diversity and randomness. The greater the randomness, the more creative it will be, but the more likely it is that false or made-up responses will appear, and yet, even in the case of deterministic decoding, the variability will be decreased, but still, no hallucinations will be completely removed.

Table 2: Categories of Hallucination in LLM Outputs

Category	Description	Example Impact
Factual Hallucination	Generation of incorrect or unverifiable factual information	Medical misinformation leading to incorrect treatment decisions
Logical Hallucination	Generation of conclusions that do not logically follow from evidence	Faulty decision support in analytics or policy recommendations
Citation Hallucination	Creation of nonexistent or inaccurate references	Misleading academic or research outputs
Contextual Hallucination	Contradiction or distortion of provided input content	Inaccurate document summarization

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

3. RETRIEVAL-AUGMENTED GENERATION (RAG): CONCEPT AND FRAMEWORK

The paradigm of Retrieval-Augmented Generation (RAG) has become a high-profile in terms of architecture intended to increase the reliability, factuality, and contextualization of Large Language Models (LLMs). In contrast to the traditional generative models, which only use knowledge captured in model parameters, RAG incorporates an external retrieval mechanism of knowledge to dynamically retrieve pertinent information in the inference process [10]. RAG can be used together with generative models to create responses with verifiable support through retrieval systems, minimizing hallucinations and enhancing reliability.

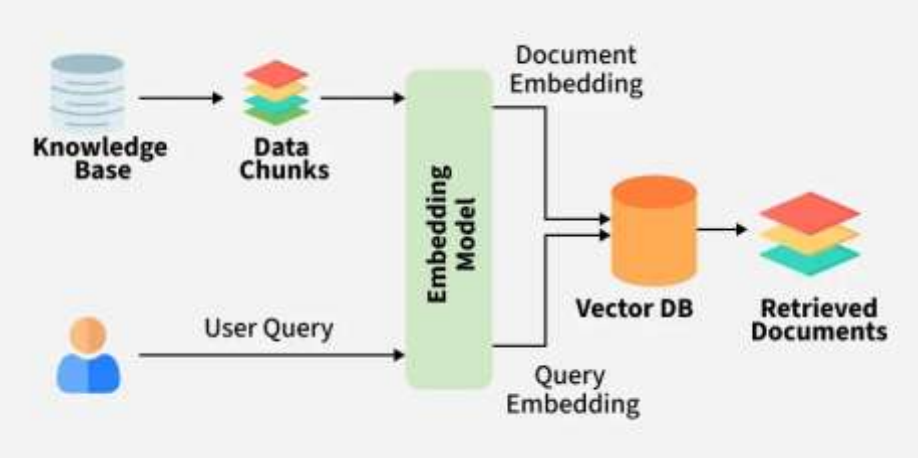


Figure 2: Architecture of Retrieval-Augmented Generation (RAG) Framework [11]

The RAG framework normally runs on a pipeline whereby a user query is handled by a retriever that indexes a knowledge base, recovers any relevant documents or passages and offers this contextual information to a generator model. The generator then generates an output contingent on the original query and retrieved evidence. Such combination of parametric knowledge (learned in model weights) and non-parametric knowledge (learned in external databases) can be seen as a major progress in the creation of explainable and updatable AI systems that are able to generate more trustworthy replies.

3.1 Architecture of RAG Systems

The RAG systems architecture consists of three major parts, the retriever, the generator, and the knowledge indexing mechanism.
Retriever Component: The role of the retriever is to locate and access pertinent information in some external knowledge repository, e.g., collections of documents, databases, or vector stores. It takes the query of the user, transforms the query into a searchable form, and ranks the candidate documents in order of their relevance [12]. The performance of the retriever plays a major role in determining the total accuracy of the generated response since the generator uses retrieved material as a contextual ground. State-of-the-art retrievers use semantic similarity and optimized indexing strategies in order to achieve high recall and precision.

Generator Component: The generator is usually a transformer based language model that generates the final response. It receives the original query and the retrieved context and generates a coherent output which incorporates the supporting evidence. The work of the generator is not merely to generate fluent text but also to make sure that the text produced by the generator is loyal to the original documents. The contemporary practice employs timely conditioning and attention strategies so that external context can be effectively added to the generation process.

Knowledge Indexing: Knowledge indexing is defined as the pre-processing and organizing external data to allow easy retrieval. The documents are divided into smaller bits, coded into searchable forms and stored in structured or better said as a vector-based database. Relevance ranking, scalability and retrieval latency are enhanced through effective indexing strategies. Periodic updates to the knowledge index can enable the RAG systems to add new or domain-specific knowledge without the need to retrain the entire model.

3.2 Types of Retrieval Techniques

RAG system performance is sensitive to the retrieval strategy used. The retrieval methods applied depend on the application needs, data type and compute limitation.

Sparse Retrieval (BM25): Sparse retrieval techniques, including BM25 algorithm, are based on term frequencyinverse document frequency (TF-IDF)-based retrieval mechanisms, and are used to match keywords between queries and documents. They are

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

computationally efficient and interpretable methods and are appropriate with structured or well-indexed textual data. Nevertheless, sparse retrieval can be weak to retrieve semantic similarity between queries and documents that have dissimilar vocabulary.

Embedding-based Retrieval: Dense retrieval applies models of neural embedding to encode query and document into continuous vectors. Cosine similarity is one of the distance measures used to compute similarity between vectors. Dense retrieval methods are useful in retrieving semantic relations, and better recall is achieved even when there are no direct matches of the query keywords. They have been popular in the current RAG implementations because of their capability to process complex and context-rich queries.

Hybrid Retrieval Approaches: Hybrid retrieval is a combination of sparse and dense retrieval methods which can use the benefits of each method. The hybrid methods enhance the precision and recall by combining keyword-based matching with semantic similarity search. Such strategies are widely used in optimized RAG frameworks in a bid to have balanced performance on various datasets and various query classes.

3.3 Benefits of RAG in Reducing Hallucination

Introduction of the retrieval mechanisms in the generation process has various strengths that directly lead to less production of hallucinations [13].

Grounded Knowledge Generation: RAG systems base generated responding on externally accessed knowledge, so that the response is based on verifiable information and not merely on the learned statistical pattern. This factual accuracy and reliability is improved.
Contextual Evidence Retrieval: RAG can also use the up-to-date domain-specific knowledge by viewing the related documents at the inference time, thereby allowing the model to use them. This minimizes the errors resulting due to outlived or unfinished pre-training data.

Explainability Via Citations: Since retrieved documents may be viewed and generated answers, RAG frameworks make it possible to have transparent and explainable outputs. The users are able to confirm the source of information, and hence they gain more confidence in the decision-making systems that are assisted by AI.

4. OPTIMIZATION STRATEGIES IN MODERN RAG FRAMEWORKS

Modern Retrieval-Augmented Generation (RAG) systems have since improved a great deal in terms of the integration of optimization methods to enhance the quality of retrieval, contextual relevance and generation fidelity [14]. As the performance of RAG is determined by the interplay between the retriever and generator units, optimization techniques are used on various pipeline stages. These techniques aim at making the recalled knowledge more accurate, making the contextual information relevant, and making the output generated more faithful. Consequently, optimized RAG models are crucial in mitigating hallucinations and enhancing the performance and reliability of responses of large language models in practice.

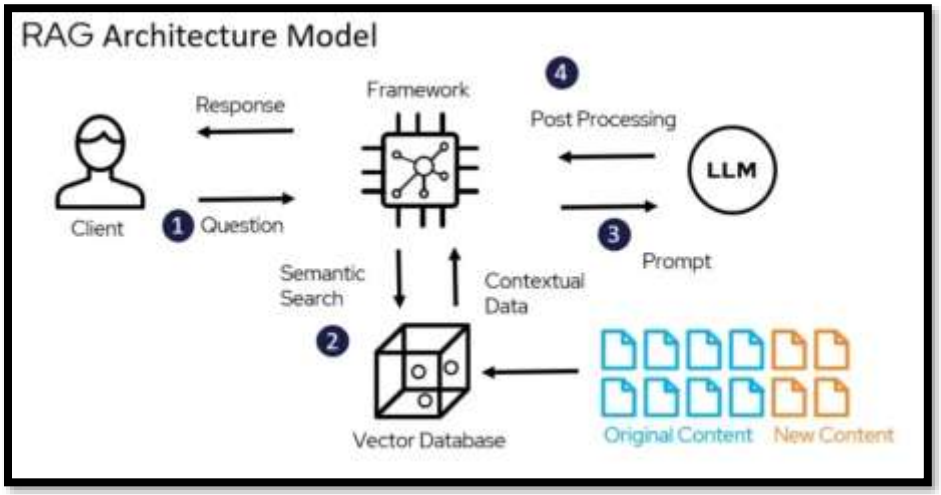


Figure 3: Optimized RAG Pipeline for Improved Retrieval and Generation Performance [15]

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

4.1 Retrieval Optimization

RAG systems depend on efficient and accurate retrieval to be successful. Enhancement at this level has a direct bearing to the quality of information availed to the generator.

Embedding Tuning: Embedding tuning: Embedding tuning is a method that refines vectors representations of queries and documents. Embedding models can be pre-trained, which does not necessarily imply excellent domain-specific semantics capturing; thus, fine-tuning embeddings on specialised data can lead to semantic similarity matching. Better embedding representations result in a high degree of relevancy during document retrieval, therefore, lessening the chances of the generator using incomplete or irrelevant context.

Query Expansion: Query expansion methods enhance the original query posted by the user by adding similar keywords, synonyms or contextual phrases [16]. This is because it enhances the recall process since the retriever is able to find the documents that might not be identical to the words in the original query. The sophisticated methods use language models to produce automatically semantically related words. Good query expansion will be able to make sure that the retrieval system traverses a wider and more pertinent evidence base, thus enhancing the accuracy of downstream generation.

Knowledge Chunking: Knowledge chunking is the division of large documents into smaller units, which are semantically relevant before indexing. Smaller chunks enhance accuracy in retrieval because they enable the system to access the most relevant passages and not full documents. Elaborate chunk size and overlap settings increase the relevance of the context and reduce noise. Scalability and lowered computational overhead on retrieval is also enhanced by optimized chunking strategies.

4.2 Reranking and Relevance Enhancement

Once document selection is performed during initial retrieval, other processes are frequently used to further narrow down document selection and pass to the generator only information that is most contextually relevant.

Cross-Encoder Reranking Cross-encoder reranking models encodes query and candidate documents together, and therefore compute relevance scores more accurately than conventional retrieval models. The cross-encoders annotate richer semantic relationships between queries and retrieved passages, though with higher computational costs, than the simpler rankers (e.g. cosine ranking) [17]. This measure would make sure that the best and most relevant documents are prioritized in generation.

Semantic Similarity Filtering: Semantic filtering algorithms are used to eliminate irrelevant or low-confidence documents in the retrieved list. Applying similarity thresholds or other techniques of clustering will maintain only extremely useful contextual evidence. The result of this filtering is a less noisy output and an avoidance of misleading or irrelevant information to the output by the generator, enhancing the faithfulness and coherence of the output.

4.3 Generation Optimization

Besides the retrieval improvements, the optimization of the generation stage is also important towards the preservation of outputs being based on retrieved evidence.

Prompt Engineering: Prompt engineering makes use of designed structured input prompts that can easily drive the generator to make use of retrieved context. Enforced format, explicit grounding criteria, and clear instructions invoke the model to give responses on the basis of available evidence and not on the basis of prior parametric knowledge. Prompts that are well-constructed can greatly decrease the amount of content that is hallucinated and also help to increase consistency in responses.

Context Window Management: Lhat language models have finite context windows, the selection and the ranking of the retrieved document have to be done efficiently [18]. The passage prioritization, summarization, and context compression techniques make sure that the most vital data is incorporated within the token constraints they have. Correct context management will eliminate loss of valuable information and improve the factual basis.

Control of Temperature and Decoding: Temperature, top-k sampling, and nucleus sampling are also generation parameters that affect variability in outputs. The low temperature conditions encourage deterministic and factually consistent answers, but those with higher temperatures are more creative, yet could introduce errors. By delicate tuning of decoding parameters, system designers can achieve the trade-off between fluency and reliability, thus lowering the incidence of hallucinations.

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

Table 3: Optimization Techniques and Performance Impact

Optimization Strategy	Purpose	Performance Benefit
Hybrid Retrieval	Improve recall by combining sparse and dense search methods	Higher factual accuracy and improved coverage of relevant information
Reranking	Improve relevance of retrieved documents	Reduced hallucination rate and enhanced contextual alignment
Prompt Optimization	Ensure effective utilization of retrieved context	More coherent, grounded, and trustworthy outputs
Embedding Tuning	Improve semantic matching between queries and documents	Increased retrieval precision and relevance
Knowledge Chunking	Enable fine-grained retrieval of relevant passages	Reduced noise and improved context utilization
Decoding Control	Manage randomness in generated outputs	Greater factual consistency and stability

5. EVALUATION METRICS AND BENCHMARKING

The analysis of Retrieval-Augmented Generation (RAG) systems is important in the context of identifying their success in minimizing hallucinations and enhancing the reliability of Large Language Model (LLM) results [19]. Because hallucination is a complex phenomenon that entails inaccuracies of facts, logical inconsistencies, and unjustified assertions, effective benchmarking entails quantitative and qualitative methods of assessment. There are modern evaluation systems which rely on automated measures, human evaluation and standard benchmark data sets to guarantee extensive performance analysis on various applications domains.

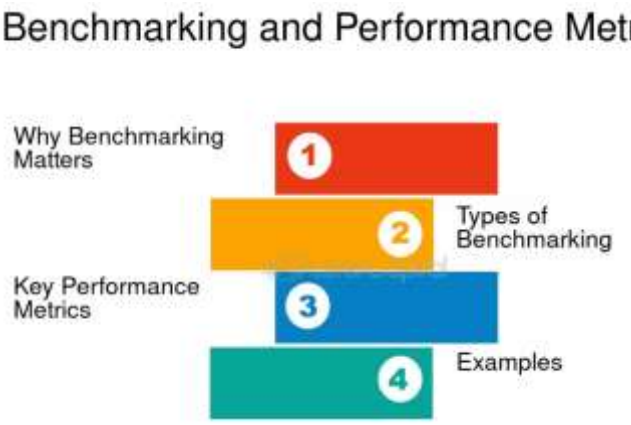


Figure 4: Conceptual Workflow of Benchmarking for System Performance Assessment [20]

5.1 Metrics for Hallucination Detection

To measure the reduction of hallucinations correctly and comprehensively, theories must have metrics that evaluate both factual accuracy and context congruence of the generated results and evidence.

Exact Match; Exact Match (EM) is a widely applicable measure in question-answering problems that analyses whether or not the response created is an exact match of the reference answer [21]. EM can also be a poor measure of model performance because semantically correct answers can be worded in different ways, which is despite the fact it offers a rigid and objective evaluation criterion. In spite of this weakness, it is useful in benchmarking the accuracy of facts in structured tasks or short-answer tasks.

Factual Consistency: Factual consistency is a metric that quantifies the consistency between generated outputs and approved sources of knowledge or retrieved evidence [22]. This measure is used to measure the support of the statements made by the model by reference documents or ground-truth datasets. Methods of measuring factual consistency are automated verification tools, entailment-based scoring, and scoring against known sources of trust. The model has high factual consistency, which means that the model has a good external information foundation.

Faithfulness; Faithfulness defines how well the generated content is related to the retrieved context without making unwarranted assertions. An accurate or faithful response is semantically consistent with the passages that have been passed on and does not create

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

new facts. Faithfulness is especially relevant in retrieval augmented systems as it quantifies solely whether or not the generator is using retrieved evidence well. Approaches to evaluation may involve human expert scoring or automated semantic similarity and entailment scoring.

5.2 Benchmark Datasets

The standardized data sets will offer a standard point of comparison of various RAG architectures and optimization strategies.
Open-Domain Question Answering Datasets; Open-domain QA datasets are popular with the aim of testing the ability to retrieve and generate language models [23]. These datasets usually have questions with a wide variety of knowledge areas, and known answers to references. These benchmarks are used to test the capacity of a model to retrieve the appropriate evidence out of a large set of corpora to create the appropriate response. They are also especially applicable to the measurement of the real-world knowledge access and performance of generalization.

Knowledge-Intensive NLP Benchmarks: Knowledge-intensive tasks involve tasks that demand more reasoning and evidence integration on the part of the model. Such benchmarks usually involve complex questions and answers, fact checking, and multi-hop reasoning problems that require any answer based on the integration of information across several sources. Testing RAG systems using these datasets will give an idea of how well they can operate with complex queries, maintain contextual grounding, and minimize hallucinated information in long-term reasoning.

5.3 Comparative Performance Analysis

Comparative performance analysis is the analysis of various RAG configurations or baseline LLMs based on standardized metrics and data. The process may involve a comparison of retrieval strategies (sparse, dense, hybrid) and reranking methods, and generation settings to which one evaluates the lowest rate of hallucination and maximum accuracy of fact. The analysis of performance scores based on various benchmarks gives an idea of optimal architectural settings and exhibits a trade off between computational efficiency and response reliability [24].

Comparative evaluation can also be used to investigate the effectiveness of individual optimization methods, such as optimization embedding tuning or adaptive retrieval, by detecting incremental performance gains in terms of accuracy, faithfulness and contextual relevance. A combination of automated measurements with human appraisal will provide a balanced measuring system performance, especially in areas that need high level of precision and interpretability.

6. CONCLUSION

This review has reviewed how the optimized Retrieval-Augmented Generation (RAG) frameworks can reduce hallucinations in Large Language Models (LLMs). Despite advanced language generation capacities, LLMs are prone to inaccuracy of information or falsification since they rely on probabilistic learning [25]. The results show that incorporation of external knowledge retrieval can greatly enhance the factual accuracy, contextual relevance and reliability of the generated outputs. The implementation of an optimization strategy that included hybrid retrieval, embedding tuning, reranking, and prompt engineering was found to be one of the major factors that decreased the cases of hallucinations. The measurement tools such as factual consistency, faithfulness and precise match also indicate quantifiable enhancement of system performance.

In spite of these developments, there are issues of scalability, maintenance of knowledge and standardized test that exist. The future studies should concentrate on adaptive retrieval, real-time integration of knowledge and improved explainability. In general, optimized RAG frameworks will be a promising solution to building trustable and evidence-based generative AI systems that can be applied to critical use.

REFERENCES

1. *Abdelghafour, M. A. M., Mabrouk, M., & Taha, Z. (2024). Hallucination mitigation techniques in large language models. International Journal of Intelligent Computing and Information Sciences, 24(4), 73-81.*
2. *Ahadian, P., & Guan, Q. (2025). A Survey on Hallucination in Large Language and Foundation Models.*
3. *Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey. arXiv preprint arXiv:2510.06265.*
4. *Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health, 4(6), e0000877.*
5. *Andriopoulos, K., & Pouwelse, J. (2023). Augmenting LLMs with Knowledge: A survey on hallucination prevention. arXiv preprint arXiv:2309.16459.*
6. *Asbai, A. (2024). Mitigating Hallucination in Large Language Model Code Generation for Higher Education: An Evaluation of Retrieval Augmented Generation.*

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 2 | February 2026 - Peer Reviewed Journal

7. *Ayala, O., & Bechard, P. (2024, June). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track) (pp. 228-238).*
8. *Brown, A., Roman, M., & Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. arXiv preprint arXiv:2508.06401.*
9. *Ding, H., Pang, L., Wei, Z., Shen, H., & Cheng, X. (2024). Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. arXiv preprint arXiv:2402.10612.*
10. *Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).*
11. *Han, B., Susnjak, T., & Mathrani, A. (2024). Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview. Applied Sciences, 14(19), 9103.*
12. *Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55.*
13. *Jiang, P., Ouyang, S., Jiao, Y., Zhong, M., Tian, R., & Han, J. (2025, August). Retrieval and structuring augmented generation with large language models. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 6032-6042).*
14. *Joshi, S. (2025). Mitigating LLM Hallucinations: A Comprehensive Review of Techniques and Architectures. Available at SSRN 5267540.*
15. *Kirchenbauer, J., & Barns, C. (2024). Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge. ArXiv Preprint. https://doi. org/10.31219/osf. io/pv7r5.*
16. *Lee, J., Cha, H., Hwangbo, Y., & Cheon, W. (2024). Enhancing large language model reliability: minimizing hallucinations with dual retrieval-augmented generation based on the latest diabetes guidelines. Journal of Personalized Medicine, 14(12), 1131.*
17. *Li, Y., Fu, X., Verma, G., Buitelaar, P., & Liu, M. (2025). Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems. arXiv preprint arXiv:2510.24476.*
18. *Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. Computers and Education: Artificial Intelligence, 100417.*
19. *Liu, S., McCoy, A. B., & Wright, A. (2025). Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. Journal of the American Medical Informatics Association, 32(4), 605-615.*
20. *Niu, M., Li, H., Shi, J., Haddadi, H., & Mo, F. (2024). Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. arXiv preprint arXiv:2405.06545.*
21. *Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. arXiv preprint arXiv:2507.18910.*
22. *Sharma, C. (2025). Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. arXiv preprint arXiv:2506.00054.*
23. *Woesle, C., Fischer-Brandies, L., & Buettner, R. (2025). A Systematic Literature Review of Hallucinations in Large Language Models. IEEE Access.*
24. *Xu, S., Yan, Z., Dai, C., & Wu, F. (2025). MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. Frontiers in Public Health, 13, 1635381.*
25. *Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics, 13(5), 856.*