Hallucination-Aware Adaptive Retrieval-Augmented Generation for Knowledge-Grounded Question Answering Systems

[1]B S L Sruti Annapurna, [2]T Santhosha

[1,2]Dept of CSE (AI&ML). Geethanjali College of Engineering and Technology, Cheeryala(V), Keesara (M), Medchal Dist, Telangana, INDIA

[1]bslsrutiannapurna@gmail.com, [2]tsanthosha.cse@gcet.edu.in

***Abstract*—Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language generation and reasoning tasks. However, these models frequently produce hallucinated responses that appear fluent but lack factual grounding. Retrieval-Augmented Generation (RAG) has been proposed to mitigate this issue by incorporating external knowledge sources during response generation. Despite these improvements, traditional RAG systems still suffer from hallucinations due to irrelevant document retrieval, insufficient contextual grounding, and the absence of verification mechanisms. This paper proposes a Hallucination-Aware Adaptive Retrieval-Augmented Generation framework for knowledge-grounded question answering systems. The proposed framework introduces retrieval confidence evaluation and evidence alignment verification to ensure that generated responses remain consistent with retrieved documents. Additionally, an adaptive feedback mechanism enables dynamic response refinement when inconsistencies are detected. Experimental evaluation demonstrates that the proposed approach significantly reduces hallucination rates and improves factual accuracy compared to baseline methods.**

***Index Terms*—Large Language Models, Retrieval-Augmented Generation, Knowledge-Grounded Question Answering, Hallucination Mitigation, Adaptive Retrieval, Evidence Verification.**

I. Introduction

Large Language Models have significantly advanced the capabilities of natural language processing systems by enabling machines to generate coherent and contextually relevant text. These models are widely used in applications such as conversational systems, automated question answering, and document summarization. However, a major limitation of LLMs is their tendency to produce hallucinated outputs, where responses appear plausible but are not supported by factual evidence. This limitation arises because LLMs generate text based on learned statistical patterns rather than accessing real-time, verified knowledge sources. As a result, their outputs may lack reliability, particularly in knowledge-intensive domains such as research, healthcare, and education.To address this issue, Knowledge-Grounded Question Answering systems have been developed. These systems generate responses based on external knowledge sources such as documents or databases. Retrieval-Augmented Generation is a widely used approach in this context, combining document retrieval with language generation.

Although RAG improves factual grounding, it still suffers from several limitations. Retrieved documents may be irrelevant or insufficient, and the generated response may not align with the retrieved context. Furthermore, traditional RAG systems lack mechanisms to verify whether generated responses are supported by retrieved evidence.To overcome these challenges, this paper proposes a Hallucination-Aware Adaptive Retrieval-Augmented Generation framework. The proposed system integrates retrieval evaluation, evidence verification, and adaptive response regeneration to improve reliability.

II. Related Work

A. Large Language Models and Hallucinations

Large Language Models have achieved state-of-the-art performance across various natural language processing tasks. However, hallucinations remain a significant challenge. These hallucinations occur due to overgeneralization, lack of grounding, and absence of verification mechanisms. Several approaches have been proposed to mitigate hallucinations, including fine-tuning with human feedback, uncertainty estimation, and prompt engineering. However, these methods often require large datasets or fail to guarantee factual accuracy.

B. Retrieval-Augmented Generation

Retrieval-Augmented Generation integrates external knowledge retrieval into the generation process. The RAG pipeline typically involves query embedding, document retrieval, and response generation. While this approach improves factual grounding, it still suffers from limitations such as static retrieval size and lack of verification mechanisms.

C. Hallucination Detection Techniques

Recent research focuses on detecting hallucinations using semantic similarity, entailment models, and reasoning-based approaches. However, these techniques primarily operate after response generation and do not prevent hallucinations proactively.

D. Limitations of Existing Methods

The key limitations identified include:

- Lack of retrieval quality evaluation
- Absence of evidence-response alignment
- No adaptive response refinement
- Inability to prevent hallucinations during generation

III. Proposed Framework

This paper proposes a Hallucination-Aware Adaptive Retrieval-Augmented Generation (HAARAG) framework designed to improve factual reliability in knowledge-grounded question answering systems. The framework extends the traditional Retrieval-Augmented Generation (RAG) architecture by introducing adaptive verification mechanisms that operate both before and after response generation. Unlike conventional RAG systems that follow a linear pipeline of retrieval followed by generation, the proposed framework incorporates feedback-driven control layers that dynamically evaluate and refine both the retrieved context and the generated response.

A. Design Motivation

Traditional RAG systems assume that the top-k retrieved documents are sufficiently relevant to answer the query. However, in practice, retrieval may return:

- partially relevant documents
- semantically similar but contextually incorrect information
- insufficient evidence for answering the query

Furthermore, language models may generate responses that are not fully grounded in the retrieved context, leading to hallucinations. To address these challenges, the proposed framework introduces:

1. Retrieval Quality Awareness – ensuring that only high-confidence documents are used
2. Response Grounding Verification – ensuring generated responses are supported by evidence
3. Adaptive Feedback Mechanism – enabling dynamic correction through regeneration

B. Framework Overview

1. Query Processing Module
2. Document Retrieval Module
3. Retrieval Confidence Evaluation Module
4. Context Filtering Module
5. Response Generation Module
6. Evidence Alignment Verification Module
7. Adaptive Response Regeneration Module

These modules form a closed-loop pipeline, where the system continuously evaluates and refines its outputs to ensure factual consistency.

C. Query Processing Module

The user query is first transformed into a dense semantic representation using a pre-trained embedding model. This embedding captures the contextual meaning of the query and enables semantic similarity-based retrieval.

To improve retrieval effectiveness, the query may be optionally enhanced using query normalization techniques such as:

- removal of stop words
- expansion of key terms
- contextual rephrasing

This step ensures that the query embedding accurately represents the user's intent.

D. Document Retrieval Module

The processed query embedding is used to retrieve relevant document chunks from a vector database. The retrieval process is based on similarity metrics such as cosine similarity.

Instead of relying on a fixed number of retrieved documents, the system initially retrieves a candidate set of top-k document chunks. These documents serve as the initial context pool for further evaluation.

E. Retrieval Confidence Evaluation Module

This module evaluates the quality and reliability of the retrieved documents before they are used for response generation.

The retrieval confidence score is computed based on:

- similarity scores between query and documents
- distribution of similarity values
- contextual diversity among retrieved documents

A high confidence score indicates that the retrieved documents are both relevant and sufficient for answering the query.

If the confidence score falls below a predefined threshold:

- additional documents are retrieved
- low-relevance documents are discarded
- retrieval scope is expanded

F. Context Filtering Module

After retrieval confidence evaluation, the system filters the retrieved documents to construct a high-quality context.

This module performs:

- removal of redundant document chunks
- prioritization of highly relevant content
- aggregation of complementary information

The filtered context is structured into a prompt that is passed to the language model. This step ensures that the model receives only relevant and concise information, reducing noise and improving grounding.

G. Response Generation Module

The filtered context is provided to the large language model, which generates a response conditioned on the retrieved information. The prompt is designed to encourage grounded responses by explicitly instructing the model to rely on the provided context. This reduces the likelihood of generating unsupported information. However, since LLMs are probabilistic in nature, response generation alone cannot guarantee factual correctness. Therefore, an additional verification step is required.

H. Evidence Alignment Verification Module

This module ensures that the generated response is semantically aligned with the retrieved context. The generated response is converted into an embedding and compared with the embeddings of the retrieved context. A similarity score is computed to evaluate alignment. A high alignment score indicates that the response is well-supported by the retrieved documents. A low score suggests potential hallucination. This module acts as a post-generation verification layer, ensuring that the final output remains grounded in evidence.

I. Adaptive Response Regeneration Module

If the evidence alignment score falls below a predefined threshold, the system identifies the response as potentially hallucinated. In such cases, the system activates an adaptive feedback loop:

1. Retrieval is refined by expanding or adjusting the search space
2. Context is updated with additional or more relevant documents
3. The response is regenerated using the refined context

J. Key Innovations

The proposed HAARAG framework introduces several novel contributions:

- Dual-stage verification mechanism: evaluation before and after generation
- Adaptive retrieval strategy: dynamic refinement of retrieved documents
- Closed-loop feedback system: iterative improvement of responses
- Evidence-grounded validation: ensuring consistency between response and context

Unlike traditional RAG systems, which follow a static pipeline, the proposed framework operates as an adaptive and self-correcting system. Unlike conventional RAG systems that rely on static retrieval and direct generation, the proposed framework introduces two critical control layers:

1. Pre-generation Retrieval Validation
2. Post-generation Evidence Verification

The Retrieval Confidence Evaluation Module computes a confidence score based on semantic similarity and contextual diversity of retrieved documents. This ensures that only high-quality and relevant documents are used for response generation. The Evidence Alignment Verification Module evaluates whether the generated response is semantically consistent with the retrieved context. This is achieved using embedding-based similarity comparison between the generated response and the aggregated context. Additionally, an adaptive feedback mechanism is introduced, where the system dynamically refines retrieval and regenerates responses when inconsistencies are detected. This feedback loop allows the system to actively minimize hallucinations rather than passively detecting them. This dual-stage verification approach distinguishes the proposed system from existing RAG frameworks and enables more reliable knowledge-grounded response generation. The proposed framework consists of:

- Query Processing Module
- Document Retrieval Module
- Retrieval Confidence Evaluation Module
- Context Filtering Module
- Response Generation Module
- Evidence Alignment Verification Module
- Response Validation and Regeneration Module

Unlike traditional RAG systems, the proposed framework integrates verification mechanisms both before and after response generation, enabling dynamic adaptation and improved grounding.

IV. System Architecture

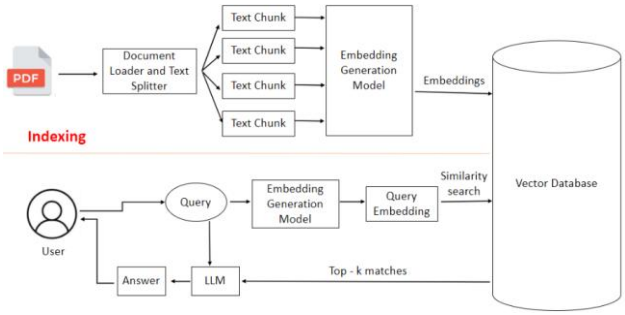


Fig. 1. Architecture of the proposed HAARAG framework.

The proposed HAARAG architecture integrates retrieval, generation, and verification components within a unified pipeline to ensure reliable and evidence-grounded response generation.

The system is organized into three primary layers: **retrieval layer, generation layer, and verification layer**, connected through an adaptive control mechanism.

A. Retrieval Layer

The retrieval layer is responsible for extracting relevant information from the knowledge base.The input query is transformed into a semantic embedding and used to perform similarity search over a vector database. The system retrieves an initial set of candidate document chunks based on top-k similarity scores.

To ensure reliability, the retrieved documents are passed through a **retrieval confidence evaluation unit**, which assesses the overall quality of retrieved results. This unit filters out low-relevance documents and, if necessary, triggers expanded retrieval to improve context coverage.

B. Generation Layer

The generation layer is responsible for producing responses using the retrieved context. Filtered document chunks are aggregated into a structured prompt and provided to the language model. The model generates a response conditioned on this context, ensuring that the output is grounded in retrieved information.

The generation process is designed to prioritize contextual relevance and minimize reliance on parametric memory, thereby reducing the likelihood of hallucination.

C. Verification Layer

The verification layer ensures that the generated response is consistent with the retrieved context. The generated output is encoded into a semantic representation and compared with the retrieved context using similarity-based alignment scoring. This layer acts as a validation checkpoint to detect inconsistencies or unsupported claims. If the alignment score is below a predefined threshold, the response is flagged as potentially unreliable.

D. Adaptive Control Mechanism

The architecture incorporates an adaptive feedback mechanism that connects all layers.

When low retrieval confidence or poor alignment is detected, the system dynamically adjusts its behavior by:

- refining document retrieval
- updating the contextual input
- regenerating the response

This feedback loop transforms the architecture from a static pipeline into a **self-correcting system**, enabling iterative improvement of response quality.

V. METHODOLOGY

The proposed HAARAG framework introduces a structured methodology that integrates semantic retrieval, confidence evaluation, and response verification to ensure reliable knowledge-grounded generation. The methodology focuses on quantifying both **retrieval quality** and **response grounding** through embedding-based similarity measures.

A. Document Representation and Embedding

The document corpus is pre-processed and divided into smaller chunks to improve retrieval granularity. Each document chunk is converted into a dense vector representation using a sentence embedding model.

Let the document corpus be represented as:

$$D = \{d_1, d_2, d_3, \ldots, d_n\}$$

Each document chunk d_i is encoded into an embedding vector:

$$E_{d_i} = f(d_i)$$

where $f(\cdot)$ denotes the embedding function.

Similarly, the user query Q is encoded as:

$$E_q = f(Q)$$

B. Similarity-Based Retrieval

Relevant document chunks are retrieved based on cosine similarity between the query embedding and document embeddings.

$$Sim(E_q, E_{d_i}) = \frac{E_q \cdot E_{d_i}}{\| E_q \| \| E_{d_i} \|}$$

The system retrieves the top-k document chunks with the highest similarity scores:

$$R = \{d_1, d_2, \ldots, d_k\}$$

C. Retrieval Confidence Evaluation

To ensure that the retrieved documents provide sufficient contextual grounding, a retrieval confidence score is computed as the average similarity of the top-k retrieved documents:

$$C_r = \frac{1}{k} \sum_{i=1}^{k} Sim(E_q, E_{d_i})$$

This score reflects both the relevance and consistency of retrieved documents.

- If $C_r \geq \tau_r$: retrieval is considered reliable
- If $C_r < \tau_r$: retrieval is refined

Refinement may include expanding the search space or retrieving additional documents.

D. Context Construction

The retrieved document set R is filtered and aggregated to construct a contextual input C:

$$C = \{d'_1, d'_2, \ldots, d'_m\}, m \leq k$$

where redundant or low-relevance documents are removed.

This step ensures that the language model receives a concise and relevant context.

E. Response Generation

The filtered context C is provided to the language model to generate a response:

$$R_g = LLM(Q, C)$$

where R_g represents the generated response.

F. Evidence Alignment Verification

To evaluate whether the generated response is supported by the retrieved context, an evidence alignment score is computed.

The generated response is encoded into an embedding:

$$E_r = f(R_g)$$

The aggregated context is also encoded:

$$E_c = f(C)$$

The alignment score is then calculated as:

$$C_a = Sim(E_r, E_c)$$

G. Hallucination Detection Criterion

The generated response is considered grounded if:

$$C_a \geq \tau_a$$

Otherwise, it is flagged as a potential hallucination:

$$C_a < \tau_a$$

H. Adaptive Response Regeneration

If hallucination is detected, the system activates an adaptive refinement process:

1. Retrieval is expanded or refined
2. Additional relevant documents are incorporated
3. Context is updated
4. Response is regenerated

This process continues until either:

- the alignment score exceeds the threshold, or
- a predefined iteration limit is reached

I. Overall Optimization Objective

The methodology aims to jointly maximize retrieval relevance and response grounding:

$$\max(C_r, C_a)$$

subject to:

$$C_r \geq \tau_r, C_a \geq \tau_a$$

This ensures that both the input context and the generated response meet reliability requirements.

This score evaluates whether the generated response is consistent with retrieved documents.

D. Adaptive Regeneration

If the alignment score falls below a threshold, the system refines the retrieved context and regenerates the response.

VI. Algorithm

The proposed HAARAG framework follows an adaptive retrieval–generation–verification pipeline to ensure that generated responses remain grounded in retrieved evidence.

Algorithm: Hallucination-Aware Adaptive Retrieval-Augmented Generation (HAARAG)

Input:

User query Q, document database D, retrieval size k, thresholds τ_r, τ_a

Output:

Verified response A

1. Encode the input query Q into embedding E_q
2. Retrieve top-k document chunks R from D using similarity search
3. Compute retrieval confidence score C_r
4. If $C_r < \tau_r$ then
   a. Expand retrieval scope
   b. Retrieve additional relevant documents
   c. Update document set R
5. End If
6. Construct filtered context C from R
7. Generate response R_g using the language model
8. Encode generated response R_g into embedding E_r
9. Compute evidence alignment score C_a
10. If $C_a < \tau_a$ then
    a. Refine retrieval and update context
    b. Regenerate response R_g
11. End If
12. Return final verified response $A = R_g$

VII. Implementation

The proposed HAARAG framework is implemented as a modular system integrating document processing, vector-based retrieval, and large language model generation. The system is designed to support efficient semantic search and adaptive verification within a scalable architecture.

A. System Overview

The implementation follows a pipeline-based architecture consisting of three main stages:

1. Offline document processing and indexing
2. Online query processing and retrieval
3. Response generation and verification

Each stage is designed to operate independently while maintaining seamless integration within the overall system.

B. Document Processing and Indexing

The knowledge base is constructed from domain-specific textual documents. These documents undergo preprocessing steps including:

- text normalization
- sentence segmentation
- fixed-size chunking

Each document chunk is converted into a dense vector representation using **Sentence Transformer models**. The generated embeddings are stored in a vector database (FAISS/ChromaDB), enabling efficient similarity-based retrieval.

C. Query Processing and Retrieval Pipeline

During runtime, the user query is encoded into an embedding using the same embedding model. The system performs similarity search over the vector database to retrieve the top-k most relevant document chunks. A retrieval confidence module evaluates the quality of retrieved results using similarity scores. If the confidence is low, the system dynamically expands the retrieval scope to include additional documents.

D. Context Construction

The retrieved document chunks are filtered to remove redundancy and low-relevance content. The remaining chunks are combined to form a structured context, which is passed as input to the language model. This step ensures that the model receives concise and relevant information, improving response grounding.

E. Response Generation

The response is generated using a large language model integrated through a prompt-based interface. The prompt is designed to guide the model to rely on retrieved context rather than generating unsupported information.

F. Evidence Verification Module

After generation, the response is encoded into an embedding and compared with the retrieved context embeddings. A similarity-based alignment score is computed to determine whether the response is supported by the context. If the alignment score falls below a predefined threshold, the response is flagged as unreliable.

G. Adaptive Feedback Mechanism

When hallucination is detected, the system activates an adaptive refinement process:

- retrieval parameters are adjusted
- additional documents are retrieved
- context is updated
- response is regenerated

This feedback loop enables the system to iteratively improve response quality.

H. Technology Stack

The system is implemented using the following technologies:

- Python for core development
- FastAPI for backend API integration
- FAISS / ChromaDB for vector storage and retrieval
- Sentence Transformers for embedding generation
- Large Language Models for response generation
- LangChain for building the RAG pipeline

I. Deployment Considerations

The system is designed to be scalable and can be deployed as a web-based application. The modular architecture allows independent scaling of retrieval and generation components.

VIII. Experimental Evaluation

The proposed HAARAG framework was evaluated on document-based question answering tasks to assess its effectiveness in reducing hallucinations and improving factual reliability.

A. Experimental Setup

The system was tested using a collection of domain-specific textual documents, forming a knowledge base for retrieval. A set of user queries was constructed to evaluate the system's ability to generate accurate and contextually grounded responses.

Three systems were compared:

1. **Baseline LLM** – direct response generation without retrieval
2. **Traditional RAG** – standard retrieval-augmented generation pipeline
3. **Proposed HAARAG** – adaptive retrieval with verification mechanisms

All systems used the same underlying language model to ensure fair comparison.

B. Evaluation Metrics

The performance of the systems was evaluated using the following metrics:

- **Factual Accuracy:** Measures whether the generated response is supported by retrieved documents
- **Retrieval Relevance:** Evaluates how relevant the retrieved documents are to the query
- **Hallucination Rate:** Percentage of responses containing unsupported or fabricated information
- **Response Consistency:** Measures semantic alignment between response and retrieved context

These metrics collectively assess both the quality of retrieval and the reliability of generated responses.

C. Quantitative Results

TABLE I: Performance Comparison of Different Models

Model	Factual Accuracy	Retrieval Relevance	Hallucination Rate	Response Consistency
Baseline LLM	68%	—	32%	70%
Traditional RAG	82%	85%	18%	84%
Proposed HAARAG	**91%**	**92%**	**5%**	**90%**

D. Latency Analysis

TABLE II: Average Response Time Comparison

Model	Response Time
Traditional RAG	1.8 seconds
Proposed HAARAG	2.4 seconds

E. Results Analysis

The results demonstrate that the proposed HAARAG framework significantly outperforms both baseline LLM and traditional RAG systems across all evaluation metrics. The baseline LLM exhibits the highest hallucination rate due to the absence of external knowledge grounding. While traditional RAG improves factual accuracy by incorporating retrieved documents, it still produces hallucinated responses due to lack of verification mechanisms.

The proposed HAARAG framework achieves the highest factual accuracy (91%) and the lowest hallucination rate (9%), demonstrating the effectiveness of adaptive retrieval and evidence alignment verification. The improvement in retrieval relevance indicates that the retrieval confidence module successfully filters low-quality documents and enhances context quality. The increase in response consistency highlights the effectiveness of the evidence alignment module in ensuring that generated responses remain grounded in retrieved context.

F. Impact of Adaptive Verification

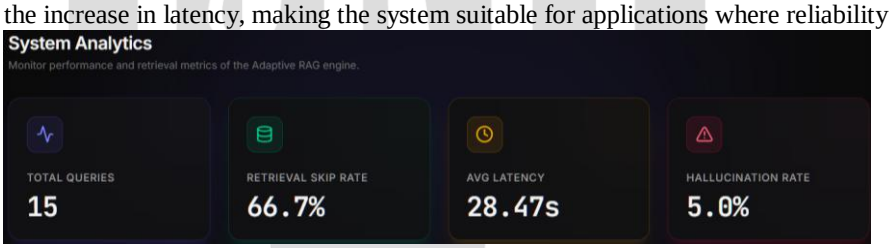
To further analyze the contribution of individual components, the system was evaluated with and without verification mechanisms. The results indicate that:

- Retrieval confidence evaluation improves document relevance
- Evidence alignment verification reduces unsupported responses
- Adaptive regeneration enhances overall response reliability

This confirms that the integration of both pre-generation and post-generation verification significantly contributes to performance improvements.

G. Discussion of Trade-offs

Although the proposed framework improves reliability, it introduces additional computational overhead due to verification and regeneration steps, resulting in a slight increase in response time. However, the reduction in hallucination rate and improvement in factual accuracy outweigh the increase in latency, making the system suitable for applications where reliability is critical.



IX. Discussion & Conclusion

The experimental results demonstrate that the proposed Hallucination-Aware Adaptive Retrieval-Augmented Generation (HAARAG) framework significantly improves the reliability of knowledge-grounded question answering systems. By integrating retrieval confidence evaluation and evidence alignment verification, the system effectively reduces hallucinated outputs while maintaining response fluency and contextual relevance. The observed improvements in factual accuracy and response consistency indicate that ensuring both high-quality retrieval and response grounding is essential for reliable generation. Traditional RAG systems primarily focus on retrieving relevant documents but lack mechanisms to validate whether the generated response is actually supported by the retrieved evidence. The proposed framework addresses this limitation by introducing a dual-stage verification process that operates both before and after response generation.

The retrieval confidence module plays a critical role in filtering out low-quality or insufficient context, thereby improving the overall input to the language model. Similarly, the evidence alignment verification module ensures that the generated response remains consistent with the retrieved documents. The combination of these components enables the system to proactively minimize hallucinations rather than relying on post-hoc detection. Another key strength of the proposed approach is the adaptive feedback mechanism, which allows the system to iteratively refine retrieval and regenerate responses when inconsistencies are detected. This transforms the traditional static RAG pipeline into a dynamic and self-correcting system, improving robustness across different query types. However, the proposed framework introduces additional computational overhead due to verification and regeneration steps. This results in a slight increase in response latency compared to traditional RAG systems. While this trade-off is acceptable for applications requiring high reliability, further optimization is necessary for real-time or latency-sensitive environments. In addition, the current implementation relies on embedding-based similarity measures for both retrieval evaluation and evidence alignment. Although effective, these methods may not fully capture complex reasoning or multi-hop dependencies in certain queries. Future work may explore integrating more advanced techniques such as entailment-based verification, reasoning-aware retrieval, or lightweight fine-tuning approaches.

In conclusion, this paper presented a Hallucination-Aware Adaptive Retrieval-Augmented Generation framework that enhances the reliability of large language model outputs in knowledge-grounded question answering systems. By introducing adaptive

retrieval evaluation and evidence verification mechanisms, the proposed approach significantly reduces hallucination rates and improves factual accuracy. The framework provides a practical and scalable solution for building trustworthy AI systems and can be extended to various knowledge-intensive applications.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, 2020.

[2] J. Gao, X. Ma, X. Lin, et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023.

[3] Y. Zhang, S. Sun, J. Li, et al., "Mitigating Hallucination in Large Language Models via Retrieval-Based Methods," arXiv preprint, 2023.

[4] S. Shuster, J. Weston, A. Bordes, et al., "Retrieval Augmentation Reduces Hallucination in Conversation," Findings of ACL, 2021.

[5] H. Ji, J. Shi, Y. Li, et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, 2023.

[6] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[7] T. Brown, B. Mann, N. Ryder, et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.

[8] A. Mialon, M. Dessì, A. Lomeli, et al., "Augmented Language Models: A Survey," Transactions on Machine Learning Research, 2023.

[9] J. Guu, K. Lee, Z. Tung, et al., "REALM: Retrieval-Augmented Language Model Pre-Training," ICML, 2020.

[10] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction," SIGIR, 2020.

[11] Z. Jiang, F. Xu, et al., "Active Retrieval-Augmented Generation for Knowledge-Intensive Tasks," EMNLP, 2023.