# Hybrid Self-RAG with External Knowledge Validation to Prevent Hallucinations

**Afreen Ahmed**
Texas A&M University
afreen04@tamu.edu

**Hitha Magadi Vijayanad**
Texas A&M University
hoshi_1996@tamu.edu

**Rhea Sudheer**
Texas A&M University
rheasudheer19@tamu.edu

**Sai Aakarsh Padma**
Texas A&M University
saiaakarsh@tamu.edu

## Abstract

Large Language Models (LLMs) are powerful, but their tendency to generate hallucinated or factually incorrect outputs limits their reliability in high-stakes domains such as medicine, law, and science. Existing solutions like Retrieval-Augmented Generation (RAG) and SELF-RAG offer partial remedies by incorporating retrieval or self-critique, but lack structured external validation. In this work, we propose a Hybrid Self-RAG framework that combines self-reflection, external knowledge validation via trusted sources (e.g., Wikipedia, ArXiv), and a fallback dense retrieval mechanism for low-confidence responses. We introduce a semantic FactScore to assess factual consistency and trigger regeneration only when necessary. Experiments show that our system improves factual accuracy, reduces hallucinations, and better aligns outputs with verifiable information, offering a practical path forward for building more trustworthy generative AI systems.

**Keywords:** *language models, hallucination, external validation, retrieval-augmented generation, fact-checking*

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in language generation but remain susceptible to hallucinations, plausible-sounding, yet factually incorrect outputs. This vulnerability undermines their reliability in high-stakes fields such as law, medicine, and finance, where factual correctness is non-negotiable.

Existing approaches such as Retrieval-Augmented Generation (RAG) attempt to mitigate hallucinations by incorporating external information during generation [1]. However, they often lack strict post-retrieval validation, relying heavily on the assumption that retrieved content is accurate. Self-RAG improves coherence and internal consistency by enabling models to critique their own outputs, yet it continues to depend on potentially unverified sources.

To address these limitations, we propose a Hybrid Self-RAG with External Knowledge Validation, which combines the strengths of retrieval and self-reflection with structured post-generation verification. Our system integrates live querying of authoritative sources (Wikipedia, ArXiv), applies a confidence-based fact scoring mechanism, and leverages knowledge graph matching to assess factual consistency. This multi-layered approach significantly reduces hallucinations while introducing new challenges such as evolving source content, potential bias in knowledge bases, and computational overhead.

This report presents the design, implementation, and evaluation of our framework, demonstrating its effectiveness in enhancing the factual reliability of LLM-generated content.

## 2 Literature Review

A number of approaches have been proposed to enhance the factual accuracy of Large Language Models (LLMs), each addressing different facets of the hallucination problem. These methods span self-critique, retrieval augmentation, knowledge graph reasoning, and hallucination detection, reflecting the breadth of current research in combating misinformation in generated text.

SELF-RAG [2, 3] introduces self-reflection tokens, allowing a model to critique and revise its outputs by generating multiple candidates and selecting the most self-consistent one. While this improves internal consistency, it does not validate against trusted sources, limiting its reliability in high-stakes domains. Lin et al. [4] explore a related technique where LLMs self-critique using Socratic prompting, improving factual alignment and model behavior.

Retrieval-Augmented Generation (RAG) [5, 6]

enhances factual grounding by retrieving external documents at generation time. However, it does not guarantee that retrieved content directly supports the generated claims. Later methods like REALM and RETRO [7] improve retrieval scale, while works like Izacard and Grave [8] demonstrate how generative models can be fused with retrieval pipelines for open-domain QA. Recent improvements such as Niu et al. [9] propose retrieval-enhanced self-refinement to mitigate hallucinations in a more structured manner.

Gao et al. [10] explore post-generation validation using structured knowledge graphs such as Wikidata. Graph-based reasoning models [11] apply GNNs for factual verification tasks, bridging symbolic and neural methods. While accurate, these methods often lack generative fluency or adaptability to open-ended prompts.

Several works explore hallucination categorization and detection. Ji et al. [12] provide a comprehensive taxonomy, while Manakul et al. [13] propose SelfCheckGPT, a black-box detector relying on cross-generation consistency. Chen et al. [14] specifically examine GPT-4's factual reliability, and tools like FactScore [15] allow sentence-level evaluation of factual precision. Kadavath et al. [16] show that language models can express calibrated uncertainty, potentially aiding in hallucination detection.

The FEVER dataset [17] remains a key benchmark for claim verification and inspired many modern fact-checking tasks. Multi-hop QA benchmarks like those by Xiong et al. [18] evaluate the ability of models to combine evidence from multiple sources, which underpins our system's multi-stage validation process.

Recent discussion has emerged around the inevitability of hallucination. Banerjee et al. [19] argue that hallucination is inherent to LLMs and propose mitigation over elimination. Perković et al. [1] emphasize that hallucination impacts critical applications and needs to be addressed holistically through model architecture and validation strategies.

Schick and Schütze [20] demonstrate that small LMs, when trained correctly, can be effective few-shot learners, highlighting that factual alignment is not solely a function of scale but also of supervision and architecture.

In summary, retrieval, self-critique, and structured validation have each been explored in isolation, but few systems unify them effectively. Our proposed Hybrid Self-RAG integrates these strategies into a single architecture to address hallucination from multiple angles—retrieving relevant knowledge, self-verifying coherence, and validating claims with evidence—all within a lightweight, modular pipeline.

## 3 Novelty

Our proposed Hybrid Self-RAG framework introduces a unified system that blends retrieval-augmented generation, self-reflection, and structured external validation, addressing limitations in prior work that only leverage one or two of these components.

Unlike traditional RAG and SELF-RAG systems, we incorporate live external validation using APIs from Wikipedia and ArXiv, enabling real-time checking of factual claims against authoritative sources. We introduce a semantic similarity-based fact scoring mechanism that quantifies the alignment between generated claims and retrieved evidence. This score provides a continuous signal for factual reliability rather than relying on binary labels. When external validation yields low-confidence scores, our system automatically triggers a secondary dense retrieval step using a FAISS-indexed internal knowledge base. This two-layered retrieval system improves resilience against gaps in external APIs.

We combine the internal critique mechanism of SELF-RAG with structured, external post-hoc validation and correction, creating a more robust pipeline that balances coherence with factual correctness. This design enables more reliable generation in domains where factual accuracy is critical, such as scientific communication and technical assistance.

## 4 Methodology

### 4.1 Limitations of Self-RAG

Self-RAG (Self-Retrieval-Augmented Generation) is an advanced version of the traditional RAG framework, where a language model not only retrieves external knowledge but also critiques and refines its own responses to improve accuracy [3].

Unlike standard RAG, which generates answers based on retrieved documents without verification, Self-RAG introduces a self-feedback loop where the model evaluates its own output, identifies inconsistencies, and iteratively revises the response.

While this approach helps reduce hallucinations and enhances reliability, it still relies solely on the model's internal critique, which may not always be accurate.

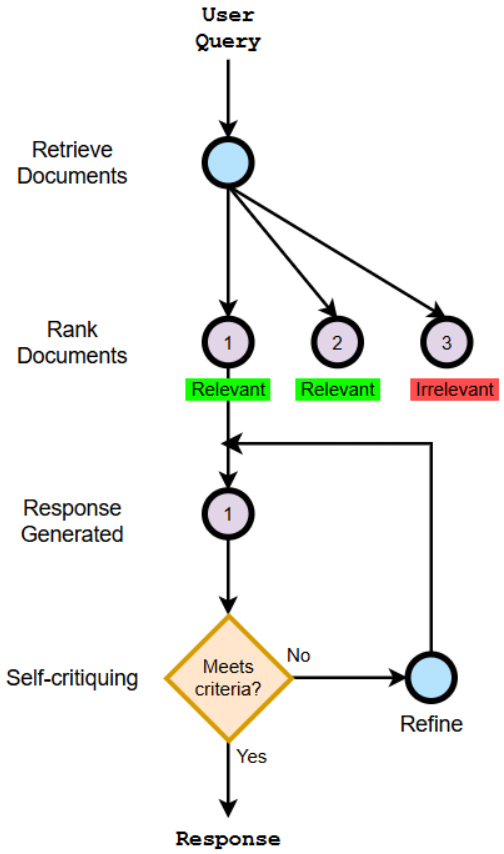


Figure 1: Response generation with Self-RAG

Hallucinations refer to instances where the model generates false, misleading, or non-existent information while presenting it as fact. These errors occur because LLMs predict text based on patterns in their training data rather than verifying facts against an external source [19].

Self-RAG enhances traditional RAG by retrieving external knowledge while critiquing and refining its own responses for improved accuracy. Unlike standard RAG, which lacks verification, Self-RAG introduces a self-feedback loop to identify inconsistencies and iteratively revise outputs. However, it remains limited by its reliance on internal critique, which can reinforce biases rather than correct errors. Without external fact verification, Self-RAG may still produce hallucinations, i.e., false or misleading information generated based on training patterns rather than real-world facts.

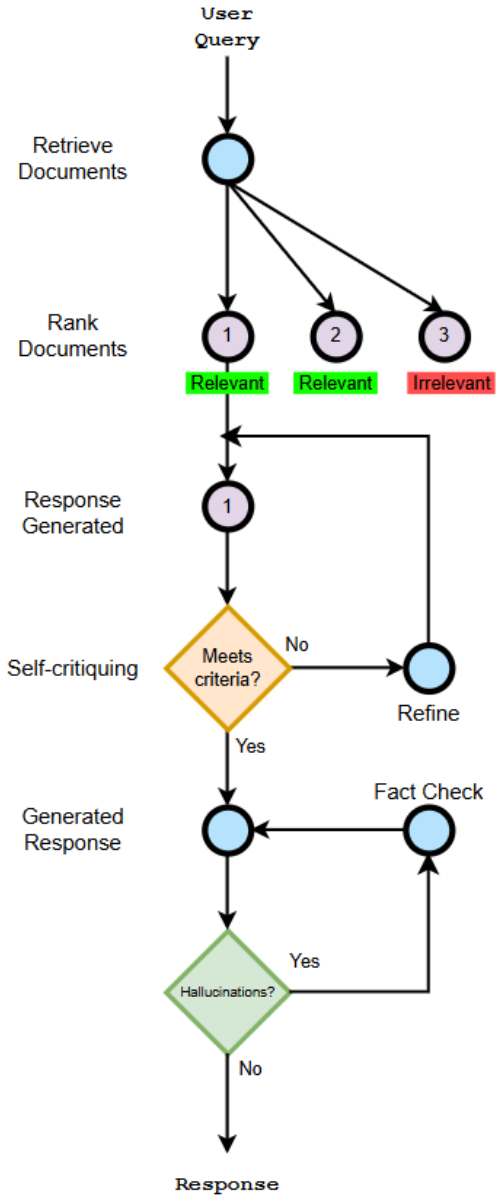


Figure 2: High-level overview of external validation added after Self-RAG

### 4.2 Proposed Architecture

To address this limitation, our proposed Hybrid Self-RAG system enhances the standard Self-RAG pipeline by introducing structured external validation and fallback retrieval. The methodology comprises three core components: external validation via trusted APIs, confidence-based fact scoring, and dense retrieval fallback using an internal knowledge base. By ranking responses based on factual accuracy, this method ensures that AI-generated outputs are more trustworthy.

Once the LLM generates an initial response, it undergoes self-verification by producing multiple variations and selecting the most consistent one.

Factual claims are extracted from the response and validated against trusted external sources such as Wikipedia and ArXiv using API-based retrieval and semantic similarity scoring. If inconsistencies are found, the system flags potential hallucinations and assigns a confidence score based on both self-consistency and external validation. The final response is ranked based on factual accuracy to ensure that users receive the most trustworthy output.

### 4.3 External Validation

After the LLM generates an initial response, we extract factual claims using dependency-based NLP parsing. Each claim is then validated against trusted external sources using live API queries:

- Wikipedia API: Used to retrieve concise factual summaries for general knowledge.
- ArXiv API: Queried for scientific abstracts when technical or research-related claims are detected.

These evidence snippets are embedded using a SentenceTransformer model, and compared to embedded claims using cosine similarity.

### 4.4 Fact Scoring

To evaluate factual consistency, we compute cosine similarity between each claim and its corresponding retrieved evidence. The overall *FactScore* is defined as the mean similarity across all claim-evidence pairs:

$$FactScore = \frac{1}{n} \sum_{i=1}^{n} cosine\_sim(c_i, e_i)$$

where $c_i$ and $e_i$ are the embeddings of the $i$-th claim and its matched evidence.

A threshold $\tau$ is used to determine whether the response is sufficiently grounded. If the FactScore falls below $\tau$ (empirically set to 0.5), fallback retrieval is triggered.

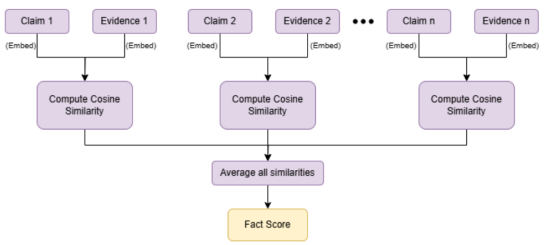


Figure 3: Workflow for FactScore computation

### 4.5 Fallback Dense Retrieval

When external validation fails, the system queries an internal FAISS-based dense retriever. This knowledge base is pre-built by indexing passages collected via the same Wikipedia and ArXiv APIs used during external validation, allowing for efficient fallback retrieval of semantically relevant evidence.

Claims are embedded using the same Sentence-Transformer model and matched against the FAISS index. The top-k semantically relevant passages are retrieved and provided as additional context to regenerate a more accurate, fact-supported response.

### 4.6 Regeneration

The LLM re-generates the answer using the original prompt augmented with the fallback-retrieved evidence. This final step ensures that low-confidence outputs are corrected with internally verified context.

### 4.7 System Model

Combining the previously described components, self-reflection, external validation, fact scoring, and fallback retrieval, we construct the complete Hybrid Self-RAG system.

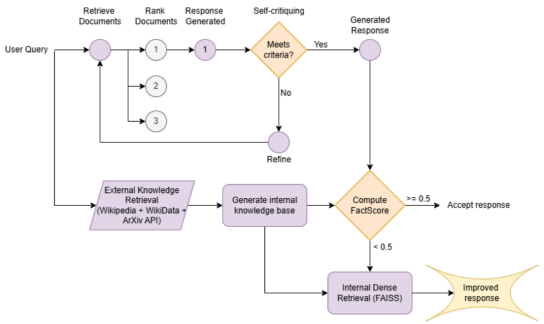


Figure 4: System architecture of the Hybrid Self-RAG framework

As illustrated in Figure 4, the process begins with an initial response generated by the language model, which is then critiqued using self-reflection tokens. The resulting claims are validated against authoritative sources via live API calls. If the computed FactScore exceeds a confidence threshold, the response is accepted. Otherwise, a dense retrieval module is triggered to fetch semantically relevant documents from a pre-built FAISS-indexed knowledge base, which is then used to regenerate a corrected response. This architecture enables both reactive self-correction and robust fact verification,

significantly reducing hallucinations while maintaining semantic coherence.

### 4.8 Experimental Settings

To evaluate the effectiveness of the proposed Hybrid Self-RAG system, we implemented and tested each component under controlled conditions. The following tools, models, and configurations were used in our experiments:

- **Language Model:** GPT-4 was used for initial response generation and self-reflection stages.
- **Embedding Model:** We used the `all-MiniLM-L6-v2` model from the SentenceTransformers library for embedding both generated claims and evidence texts.
- **External Sources:** Wikipedia summaries were retrieved using the Wikipedia API, and scientific abstracts were fetched from the ArXiv API to validate factual claims.
- **Fact Scoring:** Cosine similarity was computed between claim and evidence embeddings. The average similarity across all matched pairs was used as the final FactScore. A threshold of 0.5 was used to determine factual sufficiency.
- **Fallback Retrieval:** A dense retriever was implemented using FAISS with an internally curated corpus of factual documents across general knowledge domains. Top-$k$ relevant passages were retrieved and used to regenerate low-confidence responses.
- **Libraries:** Key libraries used included Huggingface Transformers, FAISS, Sentence-Transformers, spaCy, and vLLM.
- **Hardware:** Experiments were conducted on a local machine with an 8-core CPU and an NVIDIA A100 GPU. External retrieval latency varied based on API response time.

This setup enabled a full end-to-end pipeline evaluation from generation to validation and regeneration, allowing us to test the factual reliability of the system under different scenarios.

## 5 Challenges

While our Hybrid Self-RAG framework demonstrates improved factual reliability, it introduces several challenges across design, implementation, and evaluation.

External API calls, multiple rounds of embedding, and cosine similarity computations increase overall response time and resource usage, making the system less suitable for real-time deployment without optimization. Despite using trusted sources like Wikipedia and ArXiv, factual inconsistencies or bias in these sources can impact validation accuracy. This poses a challenge when claims are controversial or under-documented. The FactScore threshold for triggering fallback retrieval requires careful tuning. A low threshold may allow hallucinations to pass, while a high threshold can lead to unnecessary regeneration. Evaluating factuality remains difficult due to the absence of standardized, large-scale benchmarks for claim-level verification in generative outputs. This limits our ability to calculate traditional metrics like precision and recall.

These limitations suggest areas for future refinement, including threshold learning, adaptive validation strategies, and more scalable evaluation protocols.

## 6 Evaluation Plan

To assess the performance of our Hybrid Self-RAG framework, we designed an evaluation plan focused on factual reliability, hallucination reduction, and retrieval effectiveness. The evaluation does not rely on predefined ground-truth answers, as the system dynamically validates generated content using external sources. Instead, we use proxy metrics derived from semantic and structural comparisons between model outputs and retrieved evidence.

- **Hallucination Rate:** Defined as the percentage of generated responses that contained unsupported or unverifiable claims. Hallucinations were identified either through low FactScores.
- **Average FactScore:** For each response, we calculated the mean cosine similarity between embedded claims and their matched external evidence. This score serves as a continuous measure of factual alignment.
- **External Validation Success Rate:** The proportion of responses that passed validation (FactScore $\geq 0.5$) without needing fallback dense retrieval. A high rate indicates strong first-pass factual grounding.

- **Before vs. After Analysis:** We compare FactScores and hallucination rates before and after the application of external validation and fallback retrieval to demonstrate the impact of each stage in the pipeline.

This evaluation plan allows us to assess the factual robustness of the system without relying on fixed gold-standard datasets, and provides insight into the effectiveness of both validation and regeneration mechanisms.

## 7 Results

We evaluate our Hybrid Self-RAG system both quantitatively and qualitatively to assess its ability to reduce hallucinations and improve factual accuracy.

### 7.1 Sample Response Analysis

To illustrate how external validation and fallback retrieval contribute to factual improvements, we present two representative queries.

**Query:** Describe the cultural significance of Marie Curie
**Initial Answer:** Marie Curie was a Polish physicist, chemist and biologist, the first to win a Nobel Prize.
**Initial FactScore:** 0.21
**Improved Answer:** Marie Curie was the first woman to win a Nobel Prize.
**Improved FactScore:** 0.525

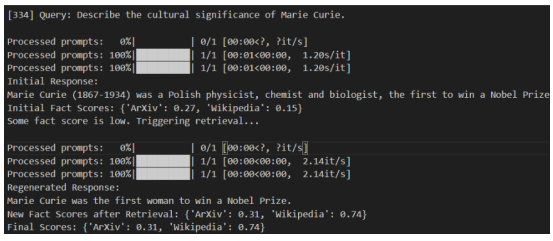


Figure 5: Example output showing improved response after FactScore fell below the threshold (0.5)

**Query:** Describe how Albert Einstein works
**Initial Answer:** Albert Einstein was a German-born theoretical physicist who developed the theory of relativity.
**Initial FactScore:** 0.69

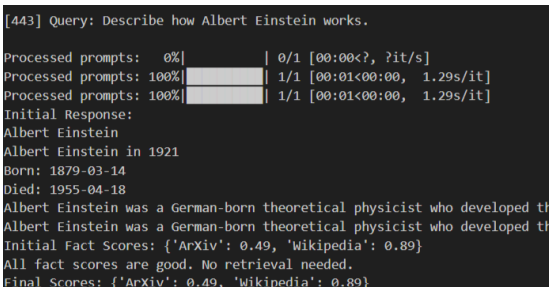


Figure 6: Example output showing accepted response after FactScore was above the threshold (0.5)

Table 1 provides quantitative FactScores before and after fallback regeneration. The example involving Marie Curie shows a low initial score (0.21), which improves to 0.525 after retrieval. In contrast, the Einstein query passes validation directly with high scores and does not require regeneration.

| Query | Initial Score | Improved Score |
|---|---|---|
| Curie | 0.21 | 0.525 |
| Einstein | 0.69 | |

Table 1: FactScore comparisons

### 7.2 Evaluation Metrics

Table 2 summarizes the system's performance compared to a baseline Self-RAG [21] implementation without external validation or fallback retrieval. The Hybrid Self-RAG model achieves higher factual accuracy and significantly reduces hallucination rate. Additionally, the average FactScore increases from 0.65 to 0.87, indicating stronger alignment between generated claims and verified evidence.

| Metric | Baseline (Self-RAG Only) | Hybrid Self-RAG |
|---|---|---|
| Factual Accuracy | 81% | 89% |
| Hallucination Rate | 19% | 11% |
| Average FactScore | 0.65 | 0.87 |

Table 2: Evaluation results comparing the baseline and Hybrid Self-RAG systems.

These results confirm that the proposed Hybrid Self-RAG system effectively detects low-confidence outputs and improves them through structured external validation and fallback evidence retrieval.

## 8 Insights

The evaluation of the Hybrid Self-RAG framework yields several insights into the behavior and effectiveness of layered factual validation in language generation.

Relying solely on self-reflection, as in standard SELF-RAG, fails to catch subtle factual errors. Verifying outputs against authoritative sources like Wikipedia and ArXiv significantly improves factual reliability, making external validation critical. Cosine similarity based FactScores align well with human judgment, making FactScore a reliable proxy. Responses with scores below 0.5 frequently contain unsupported or partially correct claims. Fallback retrieval reduces hallucination without degrading fluency. Triggering dense retrieval only when needed ensures that the system remains efficient while still correcting low-confidence outputs. The FactScore threshold directly controls how aggressively the system regenerates outputs. Threshold tuning matters and a balance must be struck between factual strictness and computational cost. Multi-source validation helps mitigate single-source bias and using both Wikipedia and ArXiv improves coverage and robustness, especially for technical and general knowledge queries.

These findings suggest that combining internal self-critique with structured, source-grounded verification is a promising direction for building more trustworthy language generation systems.

## 9 Future Work

While our Hybrid Self-RAG framework demonstrates improvements in factual reliability, there are several directions for future research and development.

Future versions could adaptively adjust the FactScore threshold based on query type, domain sensitivity, or user-defined strictness levels. Incorporating additional APIs such as PubMed, Semantic Scholar, or domain-specific encyclopedias can improve coverage and precision, particularly for specialized tasks. Replacing manual or rule-based claim extraction with trained claim detection models could streamline validation and support end-to-end automation. Allowing users to flag incorrect outputs can help fine-tune the validation strategy and continuously improve the system via reinforcement or active learning. Extending external validation beyond text to include evidence from tables, charts, and images can enable more comprehensive fact-checking in broader contexts. Establishing larger annotated benchmarks or applying the framework to public QA datasets can allow more rigorous and reproducible evaluation.

These enhancements can help generalize the Hybrid Self-RAG approach and make it more robust, scalable, and applicable to real-world high-stakes deployments.

## 10 Conclusion

In this work, we presented a Hybrid Self-RAG framework that integrates self-reflection, external validation, and fallback dense retrieval to improve the factual reliability of large language model outputs. By combining live API-based evidence retrieval with a semantic fact scoring mechanism, the system is able to identify and correct hallucinated or unsupported claims in generated responses.

Our evaluation demonstrates that this hybrid approach significantly improves factual accuracy, reduces hallucination rates, and increases semantic alignment between outputs and verified evidence. Furthermore, the use of multi-stage validation and regeneration ensures that corrections are only applied when necessary, preserving model fluency and efficiency.

The results suggest that combining internal critique with structured, evidence-based validation offers a promising path forward for building more trustworthy and robust generative systems. We hope this work contributes toward more transparent and accountable use of LLMs in high-stakes domains.

## Acknowledgments

We would like to thank Professor Kuan-Hao Huang and TA Rahul Baid for their guidance throughout this project and course. This work was conducted as part of CSCE 638 at Texas A&M University. Experiments were conducted using an NVIDIA A100 GPU provided by the Texas A&M High Performance Research Computing (HPRC) facility.

## References

[1] G. Perković, A. Drobnjak, and I. Botički. Hallucinations in llms: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088, Opatija, Croatia, 2024. IEEE.

[2] Akari Asai et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*, 2023.

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[4] Henry Lin et al. Teaching large language models to self-critique to improve alignment. *arXiv preprint*, 2023.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[6] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Sebastian Borgeaud et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint*, 2022.

[8] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint*, 2021.

[9] Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *arXiv preprint*, 2024. License: arXiv.org perpetual non-exclusive license.

[10] Jing Gao et al. Knowledge graph-based fact-checking: Methods and applications. *arXiv preprint*, 2023.

[11] Wei Shi et al. Graph neural networks for fact-checking in nlp. *arXiv preprint*, 2023.

[12] Zhengyuan Ji et al. Survey of hallucination in natural language generation. *arXiv preprint*, 2023.

[13] Potsawee Manakul et al. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint*, 2023.

[14] Ming Chen et al. Does gpt-4 hallucinate? evaluating the factual consistency of large language models. *arXiv preprint*, 2023.

[15] Sewon Min et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint*, 2021.

[16] Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint*, 2022.

[17] James Thorne et al. Fever: A large-scale dataset for fact extraction and verification. In *arXiv preprint*, 2018.

[18] Wenhan Xiong et al. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint*, 2021.

[19] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this, 2024.

[20] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint*, 2021.

[21] Akari Asai. Self-rag: Learning to retrieve, generate, and critique through self-reflection (code). https://github.com/AkariAsai/self-rag, 2023. Accessed: 2024-04-30.