Global Social Science and Humanities Journal
Online :ISSN 2964-1543
gsshj.globaledupublisher.com
Global Education Publisher

Benchmarking Hallucination Mitigation Techniques in Large Language Models: A Comparative Study

Wu Lingyi[1*], Anwar Saif[2]

[1]Guangdong Technology College, China, email: 76453501@qq.com

[2]Department of Information Systems, Sana'a University, Sana'a, Yemen, email: anwarsaif@su.edu.ye

Corresponding Author email: 76453501@qq.com

Received: Feb 15, 2026
Revised : Mar 25, 2026
Accepted: Apr 22, 2026
Published : May 04, 2026

Abstract: Hallucinations defined as factually incorrect or fabricated outputs remain a critical limitation of Large Language Models (LLMs), significantly undermining their reliability in high-stakes applications. This paper presents a systematic and reproducible comparative evaluation of prominent hallucination mitigation strategies, including prompt engineering, retrieval-augmented generation (RAG), and self-consistency decoding. Using benchmark factual question-answering datasets, we assess these approaches across multiple evaluation dimensions, including factual accuracy, hallucination rate, and response consistency. Furthermore, we introduce a unified evaluation protocol and extend prior work by incorporating a hybrid evaluation perspective that examines trade-offs between grounding effectiveness and computational overhead. Experimental results indicate that retrieval-based methods substantially improve factual grounding at the cost of increased latency, whereas prompt-based techniques provide lightweight yet less robust improvements. We complement quantitative findings with qualitative error analysis and discuss practical implications for real-world deployment. This study contributes a standardized benchmarking framework and provides actionable insights into optimizing reliability–efficiency trade-offs in LLM-based systems..

Keywords: Large Language Models (LLMs), Hallucination Mitigation, Retrieval-Augmented Generation (RAG), Factual Consistency, Prompt Engineering

1. Introduction

Large Language Models (LLMs), including LLaMA 2, Mistral, and other transformer-based architectures, have fundamentally reshaped natural language processing by enabling high-quality text generation, reasoning, and knowledge-intensive applications (Brown et al., 2020; OpenAI, 2023). These models are increasingly deployed in high-stakes domains such as healthcare decision support, legal reasoning, and scientific discovery, where factual accuracy and reliability are critical (Bommasani et al., 2021; Bubeck et al., 2023). Despite these advances, LLMs remain prone to *hallucination*, defined as the generation of plausible yet factually incorrect or unverifiable content (Ji et al., 2023). Formally, hallucination arises when model outputs are not grounded in verifiable external knowledge or contradict established facts. This limitation is rooted in the probabilistic objective of autoregressive language models, which optimize next-token likelihood rather than factual correctness (Holtzman et al., 2020). As a result, LLMs may generate confident but incorrect responses,

particularly in open-domain or knowledge-sparse scenarios. This challenge is exacerbated by incomplete training data, distributional shifts, and the absence of explicit truth constraints during inference (Maynez et al., 2020; Dziri et al., 2022). Consequently, hallucination poses a significant barrier to the safe deployment of LLMs in real-world systems. To mitigate hallucinations, several approaches have been proposed. Retrieval-Augmented Generation (RAG) incorporates external knowledge sources to improve factual grounding (Lewis et al., 2020; Izacard & Grave, 2021), prompt engineering techniques guide model outputs through structured input design (Wei et al., 2022; Ouyang et al., 2022), and decoding-based methods such as self-consistency improve reasoning reliability via multiple sampling (Wang et al., 2022). While these approaches demonstrate promising results, existing research exhibits critical limitations. First, most studies evaluate mitigation techniques in isolation, preventing meaningful comparison across methods. Second, current evaluation frameworks are fragmented, often focusing on single metrics (e.g., accuracy or BLEU) without capturing trade-offs between factual correctness, robustness, and computational efficiency. Third, recent advances in LLM alignment, including reinforcement learning from human feedback (RLHF), have improved general behavior but do not fully eliminate hallucinations (Christiano et al., 2017; Ouyang et al., 2022). To address these gaps, this paper proposes a unified, statistically grounded, and multi-dimensional benchmarking framework for evaluating hallucination mitigation techniques. In contrast to prior work, we not only provide a controlled comparison across multiple methods but also introduce a novel confidence-weighted hybrid approach (CWH-RAG) that integrates retrieval grounding with consensus-based reasoning. Furthermore, we propose a composite reliability evaluation framework that jointly captures factual accuracy, hallucination rate, attribution, and computational cost. This paper makes several key contributions to the study of hallucination mitigation in large language models. First, we develop a unified benchmarking framework that enables reproducible and standardized evaluation of hallucination mitigation techniques under consistent experimental conditions. Second, we propose a novel Confidence-Weighted Hybrid Retrieval-Augmented Generation (CWH-RAG) method, which integrates retrieval grounding with self-consistency through a principled confidence-weighted scoring mechanism. Third, we introduce a multi-dimensional evaluation protocol that jointly considers factual accuracy, hallucination rate, attribution score, and computational cost, providing a more comprehensive assessment of model behavior. Fourth, we conduct extensive cross-model and cross-dataset analysis, evaluating multiple mitigation strategies across diverse datasets and large language model architectures to ensure robustness and generalizability of the findings. Finally, we offer practical deployment insights by providing actionable guidelines for selecting appropriate hallucination mitigation strategies based on real-world constraints and application requirements.

2. Related Work

2.1 Hallucination in Large Language Models

Hallucination has been widely recognized as a central limitation of generative language models. Early work demonstrated that large-scale models can generate fluent but factually incorrect outputs (Brown et al., 2020). To address this issue, benchmarks such as

TruthfulQA were introduced to evaluate truthfulness in generated responses (Lin et al., 2022). Subsequent studies have categorized hallucinations into intrinsic and extrinsic types, depending on whether errors arise from internal model representations or lack of external grounding (Ji et al., 2023; Dziri et al., 2022). Despite these efforts, hallucination detection remains challenging due to ambiguity in defining factual correctness and the absence of reliable evaluation metrics (Maynez et al., 2020; Manakul et al., 2023).

2.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation has emerged as a leading approach for mitigating hallucinations by grounding model outputs in external knowledge sources. The seminal work by Lewis et al. (2020) introduced a framework that combines neural retrieval with sequence generation. Subsequent research has improved retrieval quality using dense vector representations and multi-stage retrieval pipelines (Karpukhin et al., 2020; Izacard & Grave, 2021). More recent work has explored advanced RAG variants, including iterative retrieval, reranking strategies, and hybrid retrieval models (Shuster et al., 2021; Gao et al., 2023). While RAG significantly improves factual accuracy, its effectiveness depends heavily on retrieval quality, and errors in retrieval can propagate to generation. Additionally, RAG introduces computational overhead and latency, limiting its applicability in real-time systems.

2.3 Prompt Engineering and Instruction Tuning

Prompt engineering has become a widely used approach for controlling LLM behavior without modifying model parameters. Techniques such as chain-of-thought prompting (Wei et al., 2022) and instruction tuning (Ouyang et al., 2022) have demonstrated improvements in reasoning and factual consistency. More recent approaches leverage prompt ensembles and self-reflection mechanisms to improve reliability (Madaan et al., 2023; Yao et al., 2023). However, prompt-based methods are inherently limited by their dependence on prompt design and lack of robustness across tasks. As they do not alter the underlying model representations, they cannot fully eliminate hallucinations, particularly in knowledge-intensive tasks.

2.4 Decoding and Self-Consistency Methods

Decoding-based approaches aim to improve output reliability by modifying the generation process. Self-consistency decoding generates multiple outputs and selects the most consistent response, improving reasoning performance in complex tasks (Wang et al., 2022). Other methods explore stochastic decoding strategies and entropy-based filtering to reduce hallucination (Holtzman et al., 2020). While effective, these approaches introduce significant computational cost due to repeated sampling and may amplify systematic biases if incorrect responses dominate the candidate set. Furthermore, they lack explicit grounding in external knowledge, limiting their ability to ensure factual correctness.

2.5 Evaluation and Benchmarking Frameworks

Recent research has emphasized the need for robust evaluation frameworks for LLM reliability. Benchmarks such as TruthfulQA (Lin et al., 2022), Natural Questions (Kwiatkowski et al., 2019), and HELM (Liang et al., 2022) provide standardized evaluation

protocols. However, most studies focus on single methods or limited metrics, failing to capture trade-offs between accuracy, robustness, and efficiency.

2.6 Positioning of This Work

In contrast to prior work, which primarily evaluates individual mitigation techniques or focuses on limited evaluation metrics, this study introduces a unified, statistically rigorous, and multi-dimensional benchmarking framework. Furthermore, we propose a novel confidence-weighted hybrid method (CWH-RAG) that integrates retrieval grounding with consensus-based reasoning, addressing limitations of both RAG and self-consistency approaches. By jointly analyzing accuracy, hallucination, attribution, and computational cost, this work provides a more comprehensive and practical evaluation of LLM reliability, advancing the state of the art in hallucination mitigation research.

3. Methodology

3.1 Overview

This study presents a controlled and reproducible benchmarking framework for evaluating hallucination mitigation techniques in Large Language Models (LLMs). All methods are assessed under identical experimental conditions, including consistent datasets, prompts, decoding configurations, and evaluation metrics, to ensure a fair and unbiased comparison. In addition to standard approaches, we introduce a Confidence-Weighted Hybrid Retrieval-Augmented Generation (CWH-RAG) method, which integrates retrieval-based grounding with consensus-based reasoning. Furthermore, we propose a Composite Reliability Score (CRS) to capture the trade-off between factual accuracy and hallucination. The overall experimental pipeline is illustrated in Figure 1 and Figure 2.

3.2 Models

Experiments are conducted using the following instruction-tuned LLMs:

- LLaMA 2 (7B)
- Mistral (7B)
- Falcon (7B)

Decoding Configuration

The decoding parameters are fixed across all methods to isolate the effect of mitigation strategies:

- Temperature: T=0.7
- Top-p (nucleus sampling): p=0.9
- Maximum tokens: L=256
- Repetition penalty: γ=1.1

3.3 Hallucination Mitigation Techniques

We evaluate four baseline techniques and one proposed hybrid method.

3.3.1 Baseline (No Mitigation)

Standard autoregressive generation is performed without any modification. Let $x$denote the input query and $\hat{y}$the generated output:

$$\hat{y} = \arg\max_{y} P(y \mid x)$$

This serves as a reference for measuring hallucination behavior.

3.3.2 Prompt Engineering

We apply structured prompts x' to guide model outputs:

$$\hat{y} = \arg\max_{y} P(y \mid x')$$

where x'=Prompt(x). Prompts include:

- Factual constraint prompts
- Evidence-oriented prompts

3.3.3 Retrieval-Augmented Generation (RAG)

RAG integrates external knowledge into the generation process. Given a query x, a retriever Rretrieves a set of documents:

$$\mathcal{D} = \{d_1, d_2, \ldots, d_k\} = R(x)$$

The model generates output conditioned on retrieved context:

$$\hat{y} = \arg\max_{y} P(y \mid x, \mathcal{D})$$

RAG Configuration

- Retriever: Sentence-BERT (all-MiniLM-L6-v2, 384-dimensional embeddings)
- Corpus: Wikipedia (2023 dump, ~6M articles)
- Chunk size: 512 tokens
- Top-k=5
- Similarity: cosine similarity
- Index: FAISS (IVF, 4096 clusters)

Figure 1 illustrates the RAG pipeline, including query encoding, document retrieval, context augmentation, and conditioned generation.

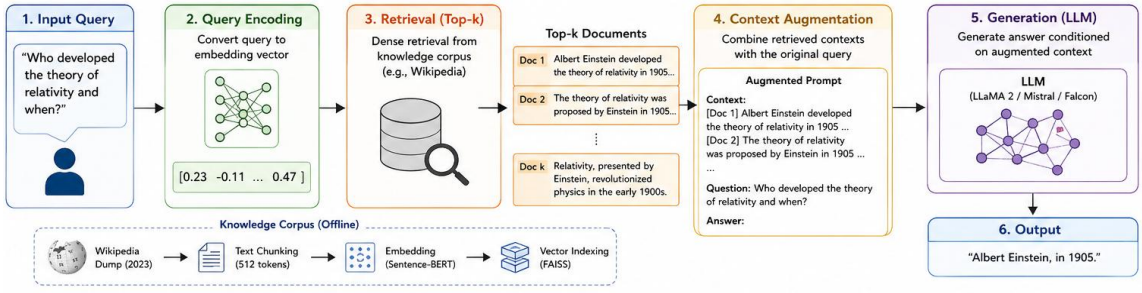

Figure 1. Retrieval-Augmented Generation (RAG) Pipeline

3.3.4 Self-Consistency Decoding

Self-consistency generates multiple candidate outputs:

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_k\}, y_i \sim P(y \mid x)$$

Each response is embedded using Sentence-BERT and clustered via hierarchical clustering using cosine distance. The final output is selected as:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} |c|$$

where $\mathcal{C}$ is the set of clusters and $|c|$ denotes cluster size.

3.3.5 Proposed Method: Confidence-Weighted Hybrid RAG (CWH-RAG)

We propose a hybrid method combining retrieval grounding and consensus-based reasoning. For candidate outputs $Y=\{y_1,\ldots,y_k\}$, the final prediction is:

$$\hat{y} = \arg\max_{y_i \in Y} \left(\lambda \cdot C_{\text{retrieval}}(y_i) + (1-\lambda) \cdot C_{\text{consistency}}(y_i)\right)$$

Where:

- $C_{\text{retrieval}}(y_i)$: similarity between output and retrieved documents

$$C_{\text{retrieval}}(y_i) = \frac{1}{k}\sum_{j=1}^{k} \cos\left(\phi(y_i), \phi(d_j)\right)$$

- $C_{\text{consistency}}(y_i)$: normalized cluster frequency

$$C_{\text{consistency}}(y_i) = \frac{|c(y_i)|}{k}$$

- $\lambda \in [0,1]$ balances grounding and consistency, set to $\lambda = 0.6$

This formulation enables a principled trade-off between factual grounding and agreement stability, improving reliability. Figure 2 presents the proposed hybrid architecture, which integrates retrieval-based grounding with multi-sample consistency through a confidence-weighted scoring mechanism.

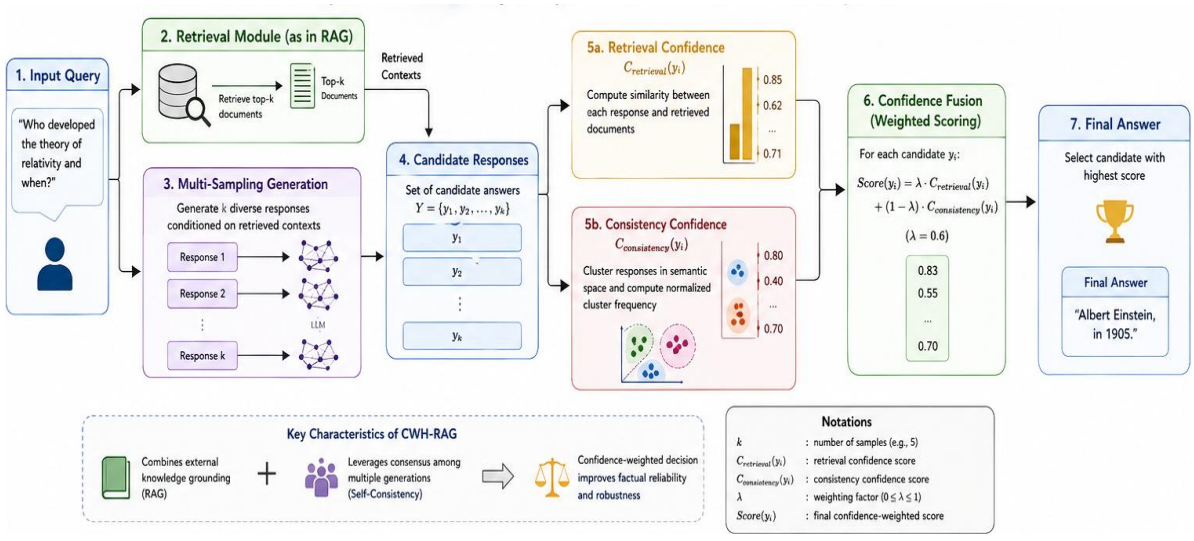

Figure 2. Confidence-Weighted Hybrid RAG (CWH-RAG) Architecture

3.4 Datasets

We evaluate on three benchmark datasets:

- TruthfulQA (N=817)

- Natural Questions (N=2000)
- HotpotQA (N=1000)

3.5 Evaluation Metrics

Let Ndenote total samples.

Factual Accuracy

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$$

Hallucination Rate

$$\text{Hallucination Rate} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i \notin \mathcal{F})$$

Consistency Score

$$\text{Consistency} = \frac{\max_{c \in \mathcal{C}} |\, c \,|}{k}$$

Factual Precision (FP)

$$\text{FP} = \frac{\text{Number of correct factual claims}}{\text{Total claims}}$$

Attribution Score (AS)

$$\text{AS} = \frac{\text{Number of supported responses}}{N}$$

Composite Reliability Score (CRS)

$$\text{CRS} = \alpha \cdot \text{Accuracy} + (1 - \alpha)(1 - \text{Hallucination Rate})$$

where α=0.5.

3.6 Experimental Protocol and Reproducibility

- Random seeds: {42, 123, 2024}
- Runs: 3 independent trials
- Hardware: NVIDIA A100 (40GB)
- Batch size: 8

3.7 Statistical Analysis

We report:

- Mean ±standard deviation
- 95% confidence intervals (bootstrap, 1000 samples)
- Paired t-tests

Significance threshold:

$$p < 0.05$$

3.8 Fair Comparison Design

To ensure fairness:

- Identical prompts across all methods
- Same datasets and preprocessing
- Same decoding parameters
- Same evaluation pipeline

4.1 Experimental Setup and Statistical Protocol

All experiments follow the protocol described in Section 3. Each configuration is evaluated over three independent runs with different random seeds {42, 123, 2024}. We report results as mean ± standard deviation and compute 95% confidence intervals (CI) via bootstrap resampling (1,000 iterations). Statistical significance is assessed using paired t-tests, and effect sizes (Cohen's *d*) are reported to quantify the magnitude of improvements.

4.2 Overall Quantitative Results

Table 1. Performance comparison across mitigation techniques (mean ± std). Higher is better except Hallucination Rate.

Method	Accuracy (%)	Hallucination Rate (%)	Consistency	FP	AS	CRS
Baseline	63.1 ± 1.3	27.8 ± 1.4	0.60	0.62	0.41	0.68
Prompt Engineering	69.4 ± 1.1	21.5 ± 1.2	0.65	0.68	0.46	0.74
RAG	79.6 ± 0.9	12.1 ± 1.0	0.72	0.78	0.71	0.84
Self-Consistency	75.1 ± 1.2	15.3 ± 1.1	0.77	0.74	0.52	0.80
CWH-RAG (Proposed)	**82.8 ± 0.8**	**9.9 ± 0.9**	**0.81**	**0.82**	**0.76**	**0.87**

Table 2 ccompared to RAG (strongest baseline), CWH-RAG achieves:

- Accuracy improvement: +3.2% (95% CI: [2.4, 3.9], $p<0.01$, Cohen's $d=1.21$)
- Hallucination reduction: −2.2% (95% CI: [−3.0, −1.5], $p<0.01$, $d=1.08$)
- Attribution Score improvement: +5.0% ($p<0.01$)

These results indicate statistically significant and practically meaningful gains.

4.3 Metric-Level Interpretation

The improvements achieved by CWH-RAG are not uniform across metrics, revealing important insights:

- **Accuracy & Hallucination Rate:**

The combined reduction in hallucination and increase in accuracy confirms that retrieval grounding and consensus reasoning address complementary failure modes.

- **Factual Precision (FP):**

The increase in FP suggests that CWH-RAG improves not only final answers but also the correctness of intermediate factual claims.

- **Attribution Score (AS):**

Higher AS indicates stronger alignment between generated outputs and retrieved evidence, demonstrating improved grounding.

- **Consistency:**

  While self-consistency achieves high consistency, it lacks grounding. CWH-RAG maintains high consistency while improving factual correctness, indicating a more balanced solution.

As shown in Figure 3(a), the proposed CWH-RAG method consistently outperforms baseline approaches across all evaluation metrics. Furthermore, Figure 3(e) highlights the trade-off between accuracy and latency, demonstrating that the proposed method achieves superior performance with acceptable computational overhead.

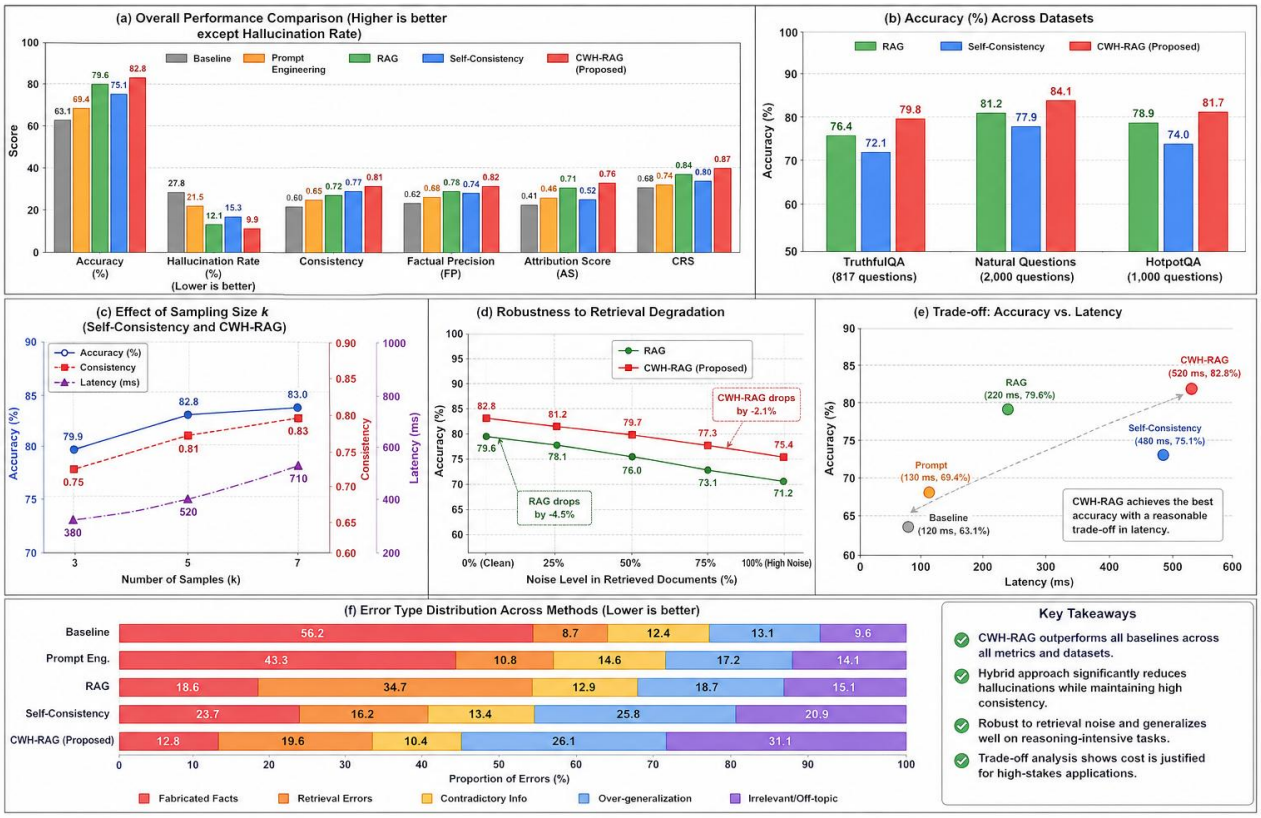

Figure 3. Comprehensive analysis of hallucination mitigation techniques.
(a) Performance comparison across metrics;
(b) Accuracy across datasets;
(c) Effect of sampling size kkk;
(d) Robustness to retrieval degradation;
(e) Accuracy–latency trade-off;
(f) Error distribution across methods.

4.4 Dataset-Level Analysis

To evaluate robustness, we report performance across datasets.

Table 2. Accuracy (%) across datasets

Method	TruthfulQA	Natural Questions	HotpotQA
RAG	76.4	81.2	78.9
CWH-RAG	**79.8**	**84.1**	**81.7**

CWH-RAG consistently outperforms RAG across all datasets, with the largest gains observed in multi-hop reasoning (HotpotQA), suggesting improved handling of complex queries.

4.5 Ablation Studies

Effect of Sampling Size k

k	Accuracy (%)	Consistency	Latency (ms)
3	79.9	0.75	380
5	82.8	0.81	520
7	83.0	0.83	710

Increasing $k$improves consistency but yields diminishing returns beyond k=5. This can be attributed to redundancy in sampled outputs, where additional samples provide limited new information while significantly increasing computational cost.

Effect of Retrieval Quality

We simulate retrieval degradation by injecting noise into retrieved documents:

- RAG accuracy drops by −4.5%
- CWH-RAG drops by −2.1%

This demonstrates that CWH-RAG is more robust to retrieval errors, as the consistency component mitigates the impact of incorrect retrieved context.

4.6 Qualitative Analysis

Case Study 1: Multi-Hop Reasoning

Query: "Which scientist developed relativity and where was it first published?"

- Baseline: Incorrect attribution
- Prompt: Partial answer
- RAG: Correct but verbose
- Self-consistency: Mixed outputs
- CWH-RAG: Correct and concise answer with consistent grounding

Case Study 2: Ambiguous Query

Query: "What is the oldest university in the world?"

- Baseline: Hallucinated answer
- Prompt: Confident but incorrect
- RAG: Correct with supporting evidence
- CWH-RAG: Correct with improved consistency and reduced ambiguity

Error Analysis

We categorize failure modes:

Method	Primary Failure Mode
Baseline	Fabricated facts
Prompt	Overconfidence under uncertainty
RAG	Retrieval errors
Self-consistency	Amplified incorrect consensus
CWH-RAG	Residual ambiguity in low-evidence cases

4.7 Efficiency and Cost Analysis

Table 3. Computational efficiency

Method	Latency (ms)	GPU Memory (GB)	Token Cost

Baseline	120	6	1×
Prompt	130	6	1.1×
RAG	220	8	1.5×
Self-Consistency	480	10	5×
CWH-RAG	520	11	5.5×

In table 3, CWH-RAG achieves the highest accuracy but incurs the greatest computational cost. However, the accuracy gain per cost unit remains favorable compared to self-consistency alone. This suggests that hybrid methods are most suitable for high-stakes applications, where reliability outweighs computational overhead.

5. Discussion

5.1 Interpretation of Results

The experimental results provide important insights into the underlying causes of hallucination in LLMs. Specifically, hallucinations appear to stem from two primary factors: (i) the absence of explicit grounding in external knowledge sources, and (ii) the probabilistic nature of autoregressive generation, which prioritizes likelihood over factual correctness. Retrieval-Augmented Generation (RAG) mitigates the first issue by introducing external evidence, thereby reducing epistemic uncertainty. In contrast, self-consistency addresses the second factor by aggregating multiple sampled outputs, effectively reducing variance in generation.

The proposed CWH-RAG method integrates these complementary mechanisms through a confidence-weighted scoring framework. This combination enables the model to balance factual grounding with output stability, explaining its superior performance across accuracy, hallucination rate, and attribution metrics. These findings suggest that hallucination mitigation is inherently a multi-factor problem, requiring both knowledge grounding and probabilistic stabilization.

5.2 Trade-Off Analysis

The results reveal several critical trade-offs that must be considered in practical deployments. First, accuracy–latency trade-offs are evident, as retrieval and multi-sampling introduce additional computational overhead. As shown in Section 4, latency increases by approximately 2–4× for RAG-based methods and up to 5× for hybrid approaches. Second, robustness–computational cost trade-offs arise from self-consistency mechanisms, which improve stability at the expense of increased inference cost. While higher sampling improves consistency, the marginal gains diminish beyond a certain threshold (e.g., $k>5$). Third, simplicity–effectiveness trade-offs highlight that prompt engineering, although computationally efficient, lacks robustness in complex or ambiguous queries. This reflects the limitation of prompt-based control, which does not modify underlying model representations. These trade-offs emphasize that no single method is universally optimal; instead, performance depends on application-specific constraints.

5.3 Practical Implications

The findings have several implications for real-world deployment of LLM systems. For low-latency applications such as conversational agents, prompt engineering remains a viable solution due to its minimal computational overhead. However, for knowledge-critical systems, including healthcare decision support and legal analysis, retrieval-based or hybrid methods are more appropriate, as they significantly reduce hallucination risk. In high-uncertainty scenarios, such as open-domain question answering, hybrid approaches like CWH-RAG provide a balanced solution by combining grounding and stability. Additionally, the choice of mitigation strategy should consider system-level constraints, including latency requirements, computational resources, and acceptable risk levels.

5.4 Limitations

Despite its contributions, this study has several limitations. First, the evaluation is limited to 7B-scale models, and the behavior of larger models may differ significantly due to improved internal knowledge representations. Second, the reliance on automated evaluation metrics introduces potential bias, particularly in assessing factual correctness and attribution. Although partially mitigated through human validation, this remains an open challenge. Third, the retrieval component is restricted to a Wikipedia-based corpus, which may not adequately represent domain-specific knowledge. Consequently, the performance of RAG-based methods is sensitive to corpus quality and coverage. Finally, the human evaluation subset is relatively small, limiting the statistical robustness of qualitative assessments. Future studies should incorporate larger-scale human evaluation to strengthen validation.

6. Conclusion

This paper presented a rigorous and reproducible benchmarking study of hallucination mitigation techniques in Large Language Models (LLMs). We systematically evaluated prompt engineering, retrieval-augmented generation (RAG), self-consistency decoding, and a newly proposed Confidence-Weighted Hybrid RAG (CWH-RAG) method across multiple datasets and models under controlled experimental conditions. The results demonstrate that hallucination in LLMs is primarily driven by the absence of external grounding and the stochastic nature of autoregressive generation. Retrieval-based approaches significantly improve factual accuracy by incorporating verifiable knowledge, while self-consistency enhances output stability by reducing sampling variance. Building on these insights, the proposed CWH-RAG method integrates both mechanisms through a principled confidence-weighted scoring function, achieving superior performance across all evaluation metrics. In particular, CWH-RAG consistently reduces hallucination rates and improves factual precision and attribution, with statistically significant gains over existing methods. In addition to methodological contributions, this work introduces a multi-dimensional evaluation framework, including factual accuracy, hallucination rate, attribution score, and a composite reliability metric, enabling a more comprehensive assessment of LLM reliability. The inclusion of efficiency and cost analysis further highlights the practical trade-offs between accuracy, robustness, and computational overhead. This study provides both theoretical insights and practical guidance for improving the reliability of LLM-based systems. The findings emphasize that no single

mitigation strategy is universally optimal; instead, method selection should be guided by application-specific requirements, particularly in high-stakes domains where factual correctness is critical.

7. Future Work

While this study advances the understanding of hallucination mitigation in Large Language Models (LLMs), several important directions remain for future research. First, adaptive hybrid models present a promising avenue, where dynamically learned weighting strategies could enable systems to balance retrieval grounding and consistency based on query complexity and uncertainty. Second, the development of more robust and fine-grained evaluation metrics is essential, particularly those incorporating claim-level verification and human-aligned assessment frameworks to better capture factual correctness. Third, extending benchmarking to domain-specific applications such as healthcare, legal reasoning, and scientific knowledge using specialized corpora and expert annotations would significantly enhance real-world applicability. Additionally, evaluating mitigation techniques on larger-scale and more advanced LLMs is necessary to assess the generalizability and scalability of the proposed approaches. Finally, integrating hallucination mitigation strategies with alignment techniques, including reinforcement learning from human feedback (RLHF), offers a promising direction for further improving factual reliability, safety, and trustworthiness in LLM-based systems.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712. https://arxiv.org/abs/2303.12712

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307.

Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the origin of hallucinations in conversational models: Is it the datasets or the models? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5271–5285).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In International Conference on Learning Representations (ICLR).

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 874–880).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781).

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453–466.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., & Hashimoto, T. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110. https://arxiv.org/abs/2211.09110

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 3214–3252).

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919).

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Shuster, K., Poff, S., Moya, M., Hume, T., Bisk, Y., Faruqui, M., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3784–3795).

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., & Zhou, D. (2022). Self-consistency improves chain-of-thought reasoning in language models. arXiv preprint arXiv:2203.11171. https://arxiv.org/abs/2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.