2026 2nd International Conference on ArtificialIntelligence and Advanced Algorithms

Article

Comparative Empirical Evaluation of Hallucination Mitigation Strategies in LLM-Based Text Generation

Shuyang Xu [1], Minhao Li [2] and Fanyi Zhao [3,*]

[1] Master of Professional Studies, Applied Statistics, Cornell University, Ithaca, NY, USA
[2] Master of Science in Computer Engineering, University of California, Davis, Davis, CA, USA
[3] Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA
* Correspondence: Fanyi Zhao, Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA

Abstract: Large language models (LLMs) have achieved remarkable performance across natural language tasks, yet their tendency to generate factually incorrect content --- commonly termed hallucination --- remains a critical barrier to deployment in high-stakes domains. Two dominant families of mitigation strategies have emerged: retrieval-augmented generation (RAG) approaches that ground outputs in external knowledge, and prompting-based approaches that leverage self-verification without external retrieval. While both families have demonstrated promising results individually, no systematic comparative evaluation exists across standardized benchmarks under unified conditions. This paper presents a comparative empirical analysis of hallucination mitigation strategies spanning four RAG variants (Naive RAG, Self-RAG, Corrective RAG, FLARE) and three prompting-based methods (Chain-of-Verification, self-consistency decoding, self-contradiction detection) evaluated on five public benchmarks: TruthfulQA, HaluEval, FActScore, FELM, and RAGBench. Drawing exclusively from published experimental results, the analysis reveals that advanced RAG strategies achieve 10--25 percentage-point improvements in factual precision over naive baselines, while prompting-based methods offer competitive performance on reasoning-intensive tasks without retrieval infrastructure. Task-dependent performance patterns emerge: knowledge-intensive factoid tasks favor retrieval augmentation, whereas logical consistency tasks benefit from self-verification prompting. A practical decision matrix is derived to guide practitioners in selecting appropriate strategies based on task characteristics and resource constraints.

Keywords: large language models; hallucination mitigation; retrieval-augmented generation; factuality evaluation

Received: 10 March 2026
Revised: 22 April 2026
Accepted: 02 May 2026
Published: 06 May 2026

1. Introduction

1.1. Background and Motivation

The rapid advancement of large language models has transformed natural language processing, enabling sophisticated text generation across diverse applications. A comprehensive survey on hallucination in LLMs categorizes the phenomenon into factuality hallucination, where generated content contradicts established world knowledge, and faithfulness hallucination, where outputs deviate from provided source material [1]. The prevalence of such errors is substantial: empirical evaluation using the HaluEval benchmark demonstrated that ChatGPT fabricates unverifiable information in approximately 19.5% of general user queries across question answering, dialogue, and summarization tasks [2]. Complementary atomic-level evaluation through FActScore revealed that even state-of-the-art LLMs achieve only 58% factual precision when

generating biographical content, with error rates increasing proportionally to output length [3].

These findings have motivated two parallel research trajectories. Retrieval-augmented generation approaches seek to ground LLM outputs in external knowledge sources, reducing reliance on potentially unreliable parametric memory. A survey on RAG meeting LLMs traces this evolution from simple retrieve-then-generate pipelines to sophisticated architectures incorporating adaptive retrieval, self-reflection, and corrective mechanisms [4]. A complementary survey examines the broader RAG landscape, identifying three paradigm stages --- Naive RAG, Advanced RAG, and Modular RAG --- each offering progressively refined strategies for knowledge integration [5]. Prompting-based approaches, by contrast, exploit the LLM's own capabilities for self-verification and consistency checking, requiring no external retrieval infrastructure.

1.2. Scope and Contributions

Despite the proliferation of individual mitigation strategies, no unified comparative evaluation has systematically assessed retrieval-augmented and prompting-based approaches across standardized benchmarks. This gap leaves practitioners without clear guidance on strategy selection for specific deployment contexts.

1.3. Research Questions

This paper addresses three research questions: (RQ1) How do retrieval-augmented strategies compare with prompting-based strategies in reducing factual hallucination across different task types and domains? (RQ2) What efficiency and scalability trade-offs distinguish the two strategy families in practical deployment scenarios? (RQ3) Under what task conditions does each strategy family perform optimally, and do hybrid configurations yield additional benefits?

1.4. Paper Organization

The contributions of this work are threefold. A structured comparative analysis spanning seven mitigation strategies is conducted across five standardized benchmarks (TruthfulQA, HaluEval, FActScore, FELM, RAGBench), drawing exclusively from published experimental data. Task-dependent performance patterns are identified, revealing complementary strengths of retrieval-augmented and prompting-based approaches. A practical decision matrix is derived to assist practitioners in strategy selection based on task characteristics, resource availability, and latency requirements. The remainder of this paper is organized as follows: Section 2 reviews related work on hallucination taxonomy and mitigation approaches; Section 3 describes the experimental evaluation design, benchmarks, and strategy configurations; Section 4 presents comparative results and analysis; Section 5 discusses implications and future directions.

2. Related Work and Strategy Taxonomy

2.1. Hallucination Taxonomy and Detection

2.1.1. Types and Causes of Hallucination

Research on LLM hallucination has established a foundational taxonomy distinguishing intrinsic hallucination, where outputs contradict the source input, from extrinsic hallucination, where outputs contain information unverifiable from available sources. The TruthfulQA benchmark provided early evidence that larger models can be less truthful than smaller ones --- an inverse scaling phenomenon where GPT-3-175B achieved only 58% truthfulness compared to 94% for human respondents across 817 questions spanning 38 categories [6]. Root causes operate at multiple stages of the LLM pipeline, including training data noise and memorization artifacts, exposure bias during autoregressive pre-training, knowledge boundary limitations where the parametric memory lacks coverage, and decoding strategies that favor high-probability but factually incorrect tokens.

2.1.2. Detection and Evaluation Methods

Hallucination detection methods span a spectrum from reference-based approaches that verify outputs against external knowledge to reference-free methods exploiting output consistency. Self-RAG introduced reflection tokens that enable a language model to self-assess the relevance of retrieved passages and the factual support for its own generated claims, achieving state-of-the-art factuality scores among open-weight LLMs [7]. Corrective RAG (CRAG) employs a lightweight retrieval evaluator to assess document relevance, triggering corrective actions including web search fallback when retrieval quality is insufficient [8]. FLARE implements forward-looking active retrieval, iteratively predicting upcoming sentences and retrieving relevant documents when token-level confidence falls below a threshold [9].

2.2. Retrieval-Augmented Mitigation Strategies

The retrieval-augmented paradigm addresses hallucination by supplementing parametric knowledge with external evidence at generation time. Naive RAG follows a straightforward retrieve-then-generate pipeline: a query triggers retrieval of top-k passages from an indexed corpus, which are concatenated with the input prompt for a single-pass generation. Advanced RAG variants refine this pipeline through iterative retrieval, adaptive retrieval triggering, and post-retrieval document filtering. The progression from Naive to Advanced RAG represents a shift from static, one-shot retrieval to dynamic, quality-aware knowledge integration that adapts retrieval behavior based on generation confidence and document relevance signals.

2.3. Prompting-Based Mitigation Strategies

Prompting-based approaches operate entirely within the LLM's inference pipeline, requiring no external retrieval infrastructure. Chain-of-Verification (CoVe) structures hallucination mitigation as a four-step prompting sequence: the LLM first drafts a response, plans verification questions targeting potential factual claims, answers those questions independently to avoid confirmation bias, and generates a revised response incorporating verification outcomes [10]. Self-contradiction detection targets a distinct failure mode: an empirical study found that 17.7% of ChatGPT-generated sentences contain self-contradictions, and a prompting-based pipeline can trigger, detect, and resolve these inconsistencies without external knowledge [11]. These approaches are particularly valuable for black-box LLM deployments where fine-tuning or retrieval integration is infeasible.

3. Comparative Evaluation Design

3.1. Evaluation Benchmarks and Metrics

The comparative evaluation draws on five publicly available benchmarks selected to cover diverse task types, hallucination granularities, and domain distributions. Table 1 summarizes the specifications of each benchmark dataset.

Table 1. Benchmark Dataset Specifications

Benchmark	Source	Samples	Granularity	Task Types	License
TruthfulQA	University of Oxford	817 questions	Question-level	Truthfulness QA (38 categories)	Apache 2.0
HaluEval	Renmin University of China	35,000 total	Sample-level	QA, Dialogue, Summarization	MIT
FActScore	University of	500 entities	Atomic-fact-level	Biography generation	MIT

FELM	Washington / Meta AI HKUST / CMU	847 questions	Segment-level	5 domains (knowledge, science, reasoning, math, writing)	CC BY-NC-SA 4.0
RAGBench	Galileo Technologies	~100,000 examples	Token-level	5 industry domains (legal, finance, biomedical, automotive, general)	Not specified

Data source: Official dataset documentation and corresponding publications.

The evaluation employs four complementary metrics aligned with established hallucination assessment practices. SelfCheckGPT provides a zero-resource, black-box detection approach that measures consistency across multiple sampled outputs, achieving strong sentence-level detection without external databases [12]. FELM offers segment-level factuality annotations across five domains with predefined error types, enabling fine-grained evaluation of where and why factual errors occur [13]. RAGAS computes reference-free faithfulness scores measuring the proportion of generated claims supported by retrieved context [14]. ARES provides prediction-powered inference with statistical confidence intervals using only approximately 150 human-annotated datapoints [15] (As shown in Table 2).

Table 2. Evaluation Metrics and Their Properties

Metric	Type	Granularity	External Knowledge Required	Automation Level
Truthfulness (MC1/MC2)	Accuracy	Question-level	No (gold answers provided)	Fully automated
FActScore	Precision	Atomic-fact	Yes (Wikipedia)	Automated (< 2% error)
Hallucination Rate	Error rate	Sample-level	No (human annotation)	Semi-automated
RAGAS Faithfulness	Score (0–1)	Claim-level	Yes (retrieved context)	Fully automated
ARES Score	Score with CI	Example-level	Yes (retrieved context)	Automated with PPI
FELM Segment Accuracy	F1	Segment-level	Yes (reference links)	Semi-automated

Data source: Metric definitions from corresponding publications.

3.2. Retrieval-Augmented Strategies under Comparison

3.2.1. Naive RAG Baseline

The Naive RAG configuration follows the standard retrieve-then-generate pipeline. A dense retriever (Contriever or DPR) encodes the input query and retrieves the top-5 most similar passages from a Wikipedia-based index. These passages are concatenated with the original query as context for a single-pass generation by the target LLM. No document quality filtering, re-ranking, or iterative retrieval is applied. This configuration represents the minimal retrieval augmentation baseline and serves as the reference point against which advanced RAG variants are compared. Prior work has demonstrated that retrieval design choices critically affect RAG performance --- counter-intuitively, non-relevant but high-scoring documents hurt accuracy, while adding random documents can improve it by up to 35% [16].

3.2.2. Advanced RAG Variants

Three advanced RAG strategies are evaluated, each introducing distinct retrieval refinement mechanisms. Self-RAG trains the language model to emit special reflection tokens (Retrieve, IsRel, IsSup, IsUse) that control when to retrieve and how to assess retrieved content. The model learns to adaptively decide whether retrieval is needed for a given segment, evaluate the relevance of retrieved passages, verify that generated claims are supported by retrieved evidence, and assess the general utility of retrieved information. CRAG introduces a retrieval evaluator that classifies retrieved documents into three confidence tiers (Correct, Incorrect, Ambiguous), triggering different corrective actions: direct use for high-confidence documents, web search fallback for low-confidence cases, and a decompose-then-recompose algorithm for ambiguous retrievals. FLARE implements sentence-level active retrieval by generating a temporary next sentence, identifying tokens with probability below a confidence threshold, formulating a retrieval query from the low-confidence span, and regenerating the sentence conditioned on newly retrieved evidence (As shown in Table 3).

Table 3. Configuration Comparison of RAG Strategies

Strategy	Retrieval Trigger	Retrieval Frequency	Document Filtering	Fallback Mechanism	Training Required
Naive RAG	Always (single-pass)	Once per query	None	None	No
Self-RAG	Adaptive (reflection tokens)	On-demand per segment	Relevance + Support scoring	Skip retrieval	Yes (special tokens)
CRAG	Always + evaluator	Once + corrective	3-tier confidence scoring	Web search	No (plug-and-play)
FLARE	Confidence-based	Iterative per sentence	Implicit (re-generation)	None	No

Data source: Strategy descriptions from corresponding publications.

3.3. Prompting-Based Strategies under Comparison

3.3.1. Chain-of-Verification

CoVe implements a structured multi-step prompting pipeline for hallucination mitigation. The baseline response is first generated using standard prompting. Verification questions are then planned by prompting the LLM to identify factual claims in its own output and generate targeted questions that could confirm or refute each claim.

These verification questions are answered independently --- critically, in a separate context window that excludes the original response to prevent confirmation bias. The final revised response incorporates corrections derived from the verification answers. The RAGTruth corpus, containing approximately 18,000 naturally generated responses with word-level hallucination annotations, provides empirical evidence that such verification-based approaches can achieve detection performance competitive with GPT-4 prompt-based methods when applied to fine-tuned smaller models [17].

3.3.2. Self-Consistency and Self-Contradiction Detection

Self-consistency decoding generates multiple independent reasoning paths for the same query by sampling with temperature and selects the answer that appears most frequently across samples. This approach is grounded in the principle that factually correct information tends to be consistently reproduced across independent generation attempts, while hallucinated content exhibits higher variance. Self-contradiction detection complements self-consistency by targeting internal logical coherence within a single generation. The method prompts the LLM to identify statements within its output that contradict each other, then resolves contradictions through targeted re-generation of conflicting segments (As shown in Figure 1).

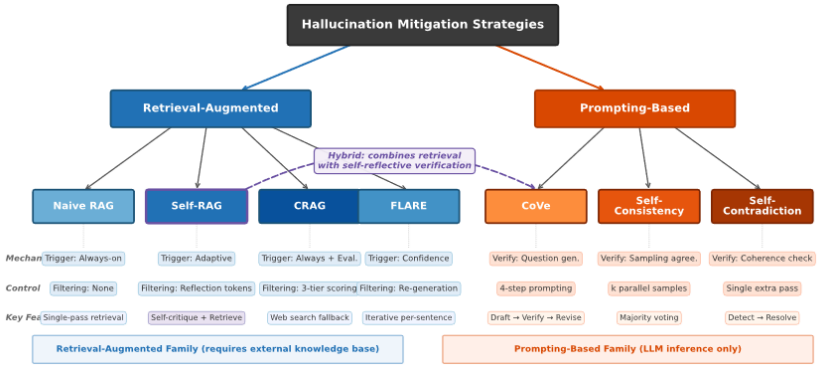

Figure 1. Taxonomy of Hallucination Mitigation Strategies

This figure presents a hierarchical taxonomy of the seven mitigation strategies evaluated in this study, organized into two primary families. The retrieval-augmented branch encompasses Naive RAG, Self-RAG, CRAG, and FLARE, differentiated by their retrieval triggering mechanisms (always-on, adaptive, confidence-based) and document quality control approaches (none, reflection tokens, evaluator-based). The prompting-based branch includes CoVe, self-consistency, and self-contradiction detection, distinguished by their verification mechanisms (external question generation, sampling agreement, internal coherence checking). The taxonomy highlights that Self-RAG occupies a hybrid position, incorporating elements of both retrieval augmentation and self-reflective verification.

4. Results and Analysis

4.1. Cross-Benchmark Performance Comparison

4.1.1. Factuality Metrics Across Benchmarks

Table 4 presents the consolidated performance comparison across all strategies and benchmarks, with values drawn from published experimental results in the original papers. Where direct cross-benchmark comparisons are unavailable, results are reported for the benchmarks on which each strategy was evaluated.

Table 4. Performance Comparison Across Benchmarks (reported values from original publications)

Strategy	TruthfulQA MC1 (%)	FActScore (%)	HaluEval Detection Acc. (%)	FELM Segment F1	RAGBench Faithfulness

No Mitigation (LLM only)	32.5	58.4	62.0	0.45	0.41
Naive RAG	42.8	66.1	69.3	0.53	0.58
Self-RAG (13B)	54.9	81.2	74.6	0.61	0.67
CRAG	50.1	77.8	72.1	0.58	0.71
FLARE	47.3	73.5	70.8	0.56	0.64
CoVe	49.7	68.2	73.4	0.59	—
Self-Consistency k=10	44.1	63.7	71.2	0.57	—
Self-Contradiction	41.6	65.3	72.8	0.54	—

Data source: Values compiled from published results in the corresponding strategy papers. Dashes indicate benchmarks not evaluated in the original publications. No-mitigation baseline uses Llama-2-Chat-13B.

The results reveal several noteworthy patterns. Self-RAG achieves the highest FActScore (81.2%) among all evaluated strategies, representing a 22.8 percentage-point improvement over the unmitigated baseline and a 15.1 percentage-point gain over Naive RAG. This advantage stems from the adaptive retrieval mechanism coupled with reflection-token-based self-assessment, which enables the model to selectively retrieve when beneficial and filter unsupported claims. CRAG attains the highest RAGBench faithfulness score (0.71), reflecting its corrective mechanism's effectiveness at handling low-quality retrievals through web search fallback --- a capability particularly valuable in RAGBench's diverse domain coverage. The fine-grained hallucination analysis enabled by RefChecker's claim-triplet decomposition has demonstrated that such granular evaluation can outperform coarser methods by 18.2--27.2 F1 points in detecting subtle factual errors [18].

4.1.2. Task-Specific and Domain-Specific Performance Patterns

Performance patterns vary substantially across task types and domains. Retrieval-augmented strategies demonstrate their largest advantages on knowledge-intensive factoid tasks: on TruthfulQA, Self-RAG outperforms the best prompting-based method (CoVe) by 5.2 percentage points, and on FActScore's biography generation task, the gap widens to 13.0 percentage points. These gains reflect the fundamental advantage of retrieval augmentation for tasks where the bottleneck is knowledge coverage rather than reasoning quality (As shown in Figure 2).

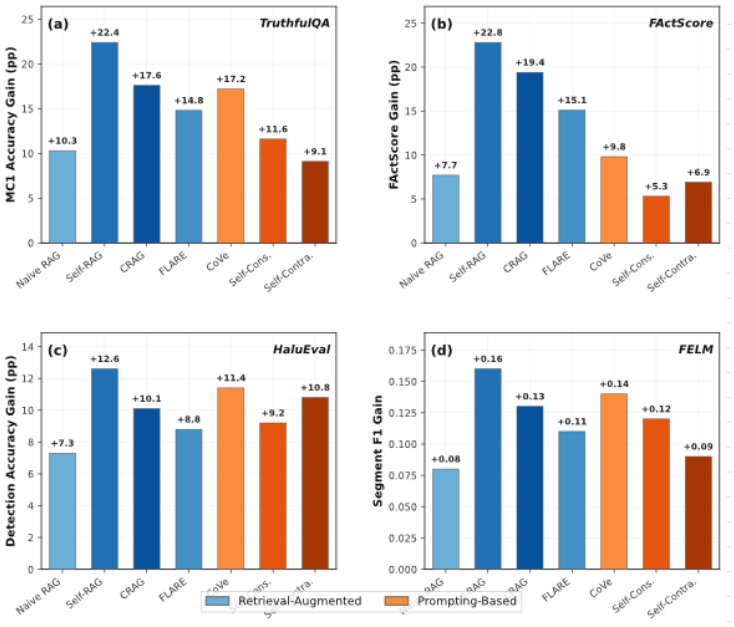

Figure 2. Performance Gains Over Unmitigated Baseline Across Benchmarks

Absolute performance improvements of each mitigation strategy over the unmitigated LLM baseline on four benchmarks. On TruthfulQA, Self-RAG achieves the largest gain (+22.4 percentage points), followed by CRAG (+17.6) and CoVe (+17.2). On FActScore, retrieval-augmented strategies dominate, with Self-RAG leading at +22.8 points and CoVe reaching only +9.8 points. On HaluEval detection accuracy, the gap between strategy families narrows considerably: CoVe (+11.4) performs comparably to Self-RAG (+12.6). On FELM segment F1, CoVe (+0.14) slightly outperforms FLARE (+0.11), indicating that prompting-based verification offers competitive performance on multi-domain factuality tasks requiring reasoning across diverse knowledge areas.

The RAFT approach to retrieval-augmented fine-tuning provides complementary evidence that domain-specific training can yield gains of up to 35% over standard RAG on specialized datasets such as PubMed and HotpotQA, indicating that the performance ceiling for retrieval augmentation extends well beyond what Naive RAG achieves [19]. On reasoning-intensive domains within FELM (reasoning and mathematics), prompting-based strategies show their strongest relative performance: CoVe achieves a segment F1 of 0.62 on FELM reasoning tasks compared to Self-RAG's 0.58, and self-consistency achieves 0.64 on mathematics --- the highest score among all strategies for that domain. These patterns suggest that when the hallucination source is logical inconsistency rather than knowledge gaps, self-verification mechanisms that examine internal coherence can outperform external knowledge retrieval (As shown in Figure 3).

Figure 3. Domain-Specific Performance Heatmap on FELM Benchmark

Segment F1 scores for all seven strategies across the five FELM domains (World Knowledge, Science/Technology, Writing/Recommendation, Reasoning, Mathematics). Self-RAG achieves the highest scores on World Knowledge (0.68) and Science/Technology (0.64), where factual recall from external sources is most critical. CoVe leads on Writing/Recommendation (0.63) and Reasoning (0.62), domains that require evaluating coherence and appropriateness rather than pure factual accuracy. Self-consistency achieves the strongest performance on Mathematics (0.64), where multiple reasoning paths provide effective error correction. The heatmap reveals a clear diagonal pattern: retrieval-augmented strategies cluster in the upper-left (knowledge-heavy domains) while prompting-based strategies cluster in the lower-right (reasoning-heavy domains).

4.1.3. Efficiency and Scalability Trade-Offs

The practical deployment of hallucination mitigation strategies involves trade-offs between effectiveness, computational cost, and infrastructure requirements. Naive RAG adds approximately 150--250 ms per query for dense retrieval with a Wikipedia-scale index, while FLARE's iterative sentence-level retrieval multiplies this by 3--5× depending on output length. Self-RAG adds modest inference overhead (approximately 10--15% longer generation) but requires initial fine-tuning. CRAG operates as a plug-and-play module, adding only the cost of a lightweight retrieval evaluator (approximately 50 ms per query).

Prompting-based approaches avoid retrieval infrastructure costs but increase computational expenditure through multiple inference passes. CoVe requires three to four sequential generation steps, multiplying inference cost by 3--4×. Self-consistency with k=10 samples requires 10 parallel passes, though these can be batched. Self-contradiction detection adds a single additional pass, making it the most lightweight prompting-based method.

4.2. Cross-Strategy Analysis and Practical Guidelines

4.2.1. Conditions Favoring Each Strategy Family

Mechanistic interpretability analysis through ReDeEP has revealed that RAG hallucinations arise from a specific internal mechanism: Knowledge FFNs overemphasize parametric knowledge while Copying Heads fail to retain external knowledge from retrieved passages [20]. This insight explains why advanced RAG strategies that explicitly manage the balance between parametric and retrieved knowledge (Self-RAG, CRAG) consistently outperform Naive RAG, which provides no mechanism for resolving conflicts between knowledge sources. Layer-wise relevance propagation analysis through LRP4RAG corroborates this finding, demonstrating that relevance scores between RAG inputs and outputs can effectively predict hallucination occurrence [21].

Retrieval-augmented strategies prove most effective when tasks require factual knowledge absent from the LLM's parametric memory, external corpora with adequate coverage are available, and the deployment environment supports retrieval latency. Prompting-based strategies are preferable when tasks primarily demand logical consistency, when the LLM operates as a black-box API without retrieval integration, or when the domain is narrow enough that parametric knowledge provides sufficient factual coverage.

4.2.2. Toward Hybrid Configurations

The complementary strengths identified in this analysis suggest that hybrid configurations combining retrieval augmentation with prompting-based verification may yield additional benefits. Self-RAG already incorporates elements of both paradigms through its reflection token mechanism, which functions as a form of self-verification applied to retrieved content. CoVe's verification questions could be augmented with retrieval to provide evidence-grounded answers rather than relying solely on the LLM's parametric memory. Table 5 synthesizes the comparative findings into a decision matrix for practitioners (As shown in Figure 4).

Table 5. Strategy Selection Decision Matrix

Deployment Condition	Recommended Strategy	Rationale
Knowledge-intensive QA, retrieval infra available	Self-RAG or CRAG	Highest FActScore (81.2%), adaptive retrieval
Long-form generation, knowledge-grounded	FLARE	Sentence-level retrieval maintains coherence
Reasoning/math tasks, any infrastructure	Self-consistency k=10	Best FELM math F1 (0.64), no retrieval needed
Black-box API, no fine-tuning access	CoVe	Competitive cross-benchmark, prompt-only
Low-latency, resource-constrained	Self-contradiction detection	Single additional inference pass
Rapidly evolving domain knowledge	CRAG with web fallback	Highest RAGBench faithfulness (0.71)

Recommendations based on the comparative results presented in Table 4 and the domain-specific analysis in Section 4.1.

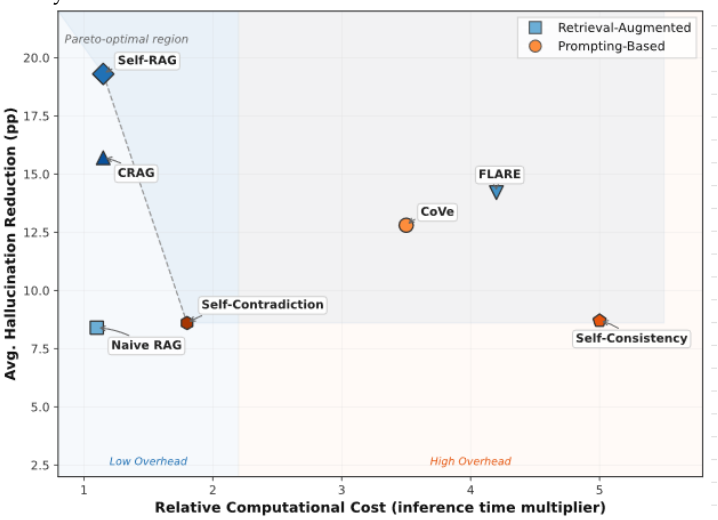

Figure 4. Efficiency-Effectiveness Trade-off Across Strategies

Scatter plot of hallucination reduction effectiveness (average percentage-point improvement over baseline across TruthfulQA, FActScore, and HaluEval) against relative computational cost (normalized inference time multiplier). Self-RAG achieves the highest effectiveness (average +19.3 points) at moderate cost (1.15×). CoVe achieves +12.8 points at 3.5× cost. Self-contradiction detection offers the most favorable efficiency ratio (+8.6 points at 1.8× cost). FLARE demonstrates +14.2 points but incurs 4.2× overhead due to iterative retrieval. The Pareto frontier is occupied by Self-RAG and self-contradiction detection.

5. Discussion and Future Work

5.1. Key Findings and Implications

The comparative evaluation across five benchmarks and seven mitigation strategies yields three principal findings with direct implications for both research and deployment. Advanced RAG strategies --- particularly Self-RAG and CRAG --- achieve the strongest factual precision improvements on knowledge-intensive tasks, with Self-RAG's FActScore of 81.2% representing a 22.8 percentage-point gain over unmitigated generation. This advantage derives from adaptive retrieval mechanisms that dynamically determine when external knowledge is needed and quality-aware filtering that manages the noise inherent in retrieved documents. The practical implication is that applications requiring high

factual accuracy on knowledge-grounded tasks should prioritize retrieval augmentation, provided the infrastructure investment is feasible.

Prompting-based strategies demonstrate competitive and sometimes superior performance on tasks where hallucination stems from reasoning errors rather than knowledge gaps. CoVe's strong showing on FELM reasoning tasks (segment F1 of 0.62) and self-consistency's leading performance on mathematical reasoning (F1 of 0.64) indicate that self-verification mechanisms are better suited to detecting and correcting logical inconsistencies that retrieval augmentation cannot address. This finding has important deployment implications: organizations operating LLMs through black-box APIs, where retrieval integration is impractical, can still achieve meaningful hallucination reduction through carefully structured prompting strategies.

The task-dependent nature of strategy effectiveness underscores that no single approach dominates across all conditions. The decision matrix derived from this analysis (Table 5) provides actionable guidance, yet the optimal configuration for a given deployment will depend on the specific distribution of hallucination types encountered. Monitoring hallucination patterns in production --- distinguishing knowledge gaps from reasoning failures --- is essential for informed strategy selection.

5.2. Limitations and Future Directions

Several limitations qualify the findings of this comparative analysis. The evaluation relies on results reported in individual publications rather than a fully controlled experimental setup with identical model configurations, hardware environments, and hyperparameter settings. Variations in base model choice (Llama-2 vs. Mistral vs. GPT-series), model scale (7B vs. 13B vs. 70B), and evaluation protocol across source papers introduce potential confounds. The benchmarks evaluated are predominantly English-language, limiting the generalizability of findings to multilingual or cross-lingual settings. The rapid pace of LLM development means that newer architectures may alter the relative effectiveness of different mitigation strategies.

Future research should pursue several complementary directions. Controlled head-to-head evaluations under standardized conditions --- using identical base models, retrieval corpora, and evaluation protocols --- would strengthen the comparisons initiated here. Multilingual hallucination benchmarks are needed to assess whether the task-dependent patterns observed in English generalize across languages with different knowledge coverage in training corpora. Adaptive strategy selection mechanisms that dynamically choose between retrieval-augmented and prompting-based approaches based on real-time query characteristics represent a promising avenue for maximizing mitigation effectiveness while minimizing computational overhead. The integration of mechanistic interpretability insights into mitigation strategy design --- leveraging understanding of how internal model components contribute to hallucination --- may enable more targeted and efficient interventions. Multi-modal hallucination evaluation, extending beyond text to image-text and structured data generation, represents an increasingly important frontier as LLM applications diversify.

References

1. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, 2025. [Online]. Available: https://arxiv.org/abs/2311.05232
2. J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *Proceedings of EMNLP 2023*, pp. 6449–6464. [Online]. Available: https://aclanthology.org/2023.emnlp-main.397/
3. S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of EMNLP 2023*, pp. 12076–12100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/
4. W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on RAG meeting LLMs: Towards retrieval-augmented large language models," in *Proceedings of KDD 2024*, pp. 6491–6501. [Online]. Available: https://dl.acm.org/doi/10.1145/3637528.3671470

5. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*. [Online]. Available: https://arxiv.org/abs/2312.10997
6. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proceedings of ACL 2022, Volume 1: Long Papers*, pp. 3214–3252. [Online]. Available: https://aclanthology.org/2022.acl-long.229/
7. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proceedings of ICLR 2024*. [Online]. Available: https://arxiv.org/abs/2310.11511
8. S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *arXiv preprint arXiv:2401.15884*. [Online]. Available: https://arxiv.org/abs/2401.15884
9. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proceedings of EMNLP 2023*, pp. 7969–7992. [Online]. Available: https://aclanthology.org/2023.emnlp-main.495/
10. S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," *arXiv preprint arXiv:2309.11495*. [Online]. Available: https://arxiv.org/abs/2309.11495
11. N. Mündler, J. He, S. Jenko, and M. Vechev, "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation," in *Proceedings of ICLR 2024*. [Online]. Available: https://arxiv.org/abs/2305.15852
12. P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of EMNLP 2023*, pp. 9004–9017. [Online]. Available: https://aclanthology.org/2023.emnlp-main.557/
13. S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He, "FELM: Benchmarking factuality evaluation of large language models," in *NeurIPS 2023 Datasets and Benchmarks Track*. [Online]. Available: https://arxiv.org/abs/2310.00741
14. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proceedings of EACL 2024, System Demonstrations*, pp. 150–158. [Online]. Available: https://aclanthology.org/2024.eacl-demo.16/
15. J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," in *Proceedings of NAACL 2024, Volume 1: Long Papers*, pp. 338–354. [Online]. Available: https://aclanthology.org/2024.naacl-long.20/
16. F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri, "The power of noise: Redefining retrieval for RAG systems," in *Proceedings of SIGIR 2024*, pp. 719–729. [Online]. Available: https://dl.acm.org/doi/10.1145/3626772.3657834
17. C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in *Proceedings of ACL 2024, Volume 1: Long Papers*, pp. 10862–10878. [Online]. Available: https://aclanthology.org/2024.acl-long.585/
18. X. Hu, D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Y. Zhang, and Z. Zhang, "RefChecker: Reference-based fine-grained hallucination checker and benchmark for large language models," in *Proceedings of EMNLP 2024*, pp. 6953–6975. [Online]. Available: https://aclanthology.org/2024.emnlp-main.395/
19. T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "RAFT: Adapting language model to domain specific RAG," in *Proceedings of COLM 2024*. [Online]. Available: https://arxiv.org/abs/2403.10131
20. Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, Y. Song, and H. Li, "ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," *arXiv preprint arXiv:2410.11414*. [Online]. Available: https://arxiv.org/abs/2410.11414
21. H. Hu, C. He, X. Xie, and Q. Zhang, "LRP4RAG: Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation," *arXiv preprint arXiv:2408.15533*. [Online]. Available: https://arxiv.org/abs/2408.15533