Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35629

A Review of Hallucination Suppression Technologies for Large Language Models Under RAG Architecture

Shujing Liu

School of International Engineering, Xi'an University of Technology, Xi'an, China
18091338851@163.com

Abstract. Large language models (LLMs) suffer inherent factual hallucination defects, which block their deployment in high-risk fields such as medicine and finance. Retrieval-Augmented Generation (RAG) serves a mainstream hallucination mitigation solution by introducing traceable external knowledge evidence. Nevertheless, existing RAG variants are plagued by retrieval noise, poor domain generalization, lack of reasoning verification and inconsistent evaluation standards. This paper adopts classification and comparative analysis as core research methods, and systematically sorts out all hallucination suppression technical routes centered on mitigating LLM hallucinations. Four major categories of anti-hallucination RAG technologies are summarized and their applicable boundaries are compared; two mainstream evaluation benchmarks, CRAG and RAGEval, are thoroughly analyzed. Aggregated experimental results demonstrate that layered stacking of multiple technologies achieves optimal hallucination reduction performance. Finally, this paper summarizes existing research gaps, including lightweight deployment and multimodal expansion, and proposes future research directions for trustworthy RAG systems. This review provides systematic theoretical support for industrial RAG model selection and optimization.

Keywords: Retrieval-Augmented Generation, Hallucination Mitigation, Large Language Model, Domain Adaptation, RAG Evaluation

1. Introduction

Powerful large language models deliver fluent text generation, yet they frequently produce fabricated, missing or contradictory factual content defined as hallucinations. Such critical flaws make LLMs unreliable for high-stakes vertical domains, including clinical diagnosis, financial analysis and legal consultation. Retrieval-Augmented Generation (RAG) becomes the dominant technical paradigm to alleviate hallucinations by attaching external retrieved factual passages to model input, grounding model outputs on real-world documents instead of parametric memory only. However, vanilla RAG with fixed top-k retrieval and frozen encoders cannot fully eliminate hallucination risks, triggering abundant improved RAG architectures proposed from 2023 to 2026.

Representative optimized RAG frameworks cover four technical branches: retrieval-layer optimized Blended RAG [1], domain-adaptive end-to-end trained RAG-end2end [2], self-reflection based Self-RAG [3], knowledge graph augmented KRAGEN [4], as well as long-context RAG anti-

DOI: 10.54254/2755-2721/2026.GL35629

noise strategies [5]. Meanwhile, two authoritative RAG evaluation benchmarks, CRAG and RAGEval, are developed to quantify hallucination severity and overall system reliability. Most existing academic papers only focus on single algorithm innovation, lacking unified horizontal comparison of all anti-hallucination pipelines under the core theme of hallucination suppression.

This paper first concludes the unified root causes of hallucinations inside standard RAG systems, then divides existing anti-hallucination approaches into four systematic technical categories: retrieval noise reduction, domain adaptive joint training, self-critical generation mechanism and knowledge graph enhanced RAG. The paper further compares the strengths, weaknesses and applicable scenarios of CRAG and RAGEval benchmarks, summarizes universal industrial architectures that stack multiple anti-hallucination modules, and identifies shared research limitations across current literature. This review distinguishes applicable boundaries of diverse anti-hallucination methods for engineering practitioners, clarifies blank research areas in the field, and guides subsequent exploration on trustworthy retrieval-augmented large language models.

2. Unified root causes of hallucinations in standard RAG

First, poor retrieval produces hard negatives: sparse/dense retrievers return semantically similar but incorrect passages. Larger Top-K improves recall yet hurts precision, confusing LLMs into contradictory responses. Long-context LLMs follow an inverted U-shaped performance curve with growing retrieved texts, especially under strong embedding retrievers [6]. Random negative benchmarks cannot replicate real hard negative interference.

Second, separate retriever-generator training weakens semantic matching. Standard RAG only fine-tunes question encoders while locking passage encoders and knowledge indexes. Without joint optimization, queries and domain documents have mismatched semantic spaces [7], leading to irrelevant retrieved content and generative hallucinations [2].

Third, LLMs lack native self-verification. Conventional RAG fetches fixed passages without dynamic retrieval judgment or evidence checking, leading to ungrounded claims and high hallucination rates in long reasoning tasks [3].

Fourth, general knowledge bases suffer domain mismatch. Wikipedia-pre-trained vanilla RAG poorly adapts to medical, financial and other vertical fields. Embedding domain gaps increase retrieval errors; static universal knowledge misses real-time, long-tail and multi-hop facts, the core source of professional QA factual mistakes [8].

3. Four main technical systems for hallucination suppression

3.1. Retrieval layer noise reduction technology

Retrieval noise reduction targets hard negative samples and low-precision retrieval results, containing two mainstream technical paths: hybrid multi-index blended RAG and long-context passage reordering strategy.

3.1.1. Blended RAG hybrid retrieval architecture

Blended RAG combines three complementary retrieval indexes: BM25 sparse full-text index, sentence transformer-based dense vector index, and Sparse EncodeR semantic sparse index. It integrates four hybrid query strategies (Cross Fields, Most Fields, Best Fields, Phrase Prefix) to unify keyword matching and semantic similarity. The framework first performs BM25 matching,

DOI: 10.54254/2755-2721/2026.GL35629

conducts hybrid retrieval over dense and sparse indexes, and selects six efficient hybrid combinations for generation.

Experiments on NQ, TREC-COVID and HotPotQA validate that the Sparse EncodeR + Best Fields hybrid query achieves Top10 retrieval accuracy of 88.77% on NQ and 98% on TREC-COVID. On SQuAD, zero-fine-tuning Blended RAG obtains an F1 score of 68.4, substantially exceeding vanilla RAG (39.42) and end-to-end RAG (52.63) with a 50% F1 improvement. In terms of storage, the dense index for the 5M HotPotQA corpus requires 50GB, while the sparse index only needs 10.5GB, offering better adaptability for large-scale enterprise corpora without metadata. Sparse EncodeR also outperforms BM25 and dense retrieval in metadata-free domain scenarios.

3.1.2. Long context retrieval reordering optimization

To mitigate the hard negative-triggered inverted U-shaped performance curve of RAG, this paper proposes a zero-overhead retrieval reordering method leveraging LLMs' "lost-in-the-middle" trait. It arranges high-relevance texts at context head and tail and low-scoring hard negatives in the neglected middle segment. This method brings slight gains under small Top-K settings yet steady prominent gains with massive retrieved passages, and it can plug into any RAG system without extra inference cost.

Two supplementary fine-tuning approaches further enhance hard negative resistance. Implicit RAG fine-tuning trains LLMs on noisy retrieved samples to master implicit noise filtering. Explicit intermediate reasoning fine-tuning adds a document relevance assessment step prior to response generation, surpassing implicit fine-tuning on multi-hop reasoning benchmarks. Cost comparison shows retrieval reordering has zero runtime cost, cross-encoder reranking brings severe latency, and RAG fine-tuning only bears one-time training expense with no inference overhead. The three mutually independent optimizations can be stacked for stronger anti-hallucination effects.

3.2. Domain adaptive training technology (RAG-end-to-end)

Original RAG freezes passage encoder and knowledge base index during fine-tuning, which severely damages cross-domain generalization. RAG-end-to-end realizes full joint training of DPR retriever and BART generator via asynchronous re-encoding and re-indexing pipelines, which forms the core domain-adaptive anti-hallucination solution.

3.2.1. Asynchronous full-link training mechanism

RAG-end-to-end divides training into three parallel processes: the main gradient update loop, multi-GPU asynchronous passage re-encoding, and FAISS asynchronous re-indexing. Re-encoding and re-indexing run parallel to training without blocking iteration, cyclically refreshing domain document embeddings to maintain up-to-date vector indexes. The framework avoids the repeated huge computation overhead of re-encoding millions of documents in each training epoch.

3.2.2. Auxiliary statement reconstruction loss

An auxiliary reconstruction training signal is introduced to strengthen domain semantic learning. A special token <p> distinguishes reconstruction input from standard QA input:

QA INPUT: ⟨Question⟩+ ⟨Retrieved Passages⟩

RECONSTRUCTION INPUT: ⟨<p>⟩+ ⟨Retrieved Passages⟩

DOI: 10.54254/2755-2721/2026.GL35629

The auxiliary task forces the model to reconstruct complete domain sentences only from retrieved chunks, strengthening the matching between query semantics and domain document features.

3.2.3. Experimental validation across three domains

Three vertical datasets (COVID-19 medical, news, conversation) are used for evaluation with five model variants compared: vanilla zero-shot RAG, QA-only fine-tuned RAG, end-to-end RAG without auxiliary loss, and end-to-end RAG with reconstruction signal [2]. RAG-end-to-end+R attains the best EM, F1, Top-5 and Top-20 retrieval scores across all domains. In conversation tasks, it lifts EM by 13 points and Top-5 retrieval accuracy by 27 points over vanilla domain-finetuned RAG. Joint end-to-end training outperforms separately optimized DPR retrievers, proving retriever-generator co-optimization critical to reduce domain hallucinations. On Wikipedia-derived SQuAD, RAG-end-to-end also beats vanilla RAG greatly (EM: 40.02 vs 28.12). Experiments further show static frozen embeddings like RETRO cannot fit vertical domains, and domain-adaptive retrievers are necessary for stable domain QA.

3.3. Self-critical generation technology (SELF-RAG)

SELF-RAG introduces reflection tokens to realize on-demand retrieval and segment-level self-critique during generation, eliminating unconditional fixed retrieval and missing evidence judgment in traditional RAG, which fundamentally cuts unsupported hallucinated content.

3.3.1. Four categories of reflection tokens

Retrieve token: Judge whether external document retrieval is required before generating each text segment [3];

ISREL token: Evaluate the relevance between the retrieved passage and the user query;

ISSUP token: Judge if the generated segment is fully supported by retrieved evidence (Fully supported / Partially supported / No support / Contradictory);

ISUSE token: Score the overall usefulness of generated answers.

3.3.2. Two-stage training pipeline

Stage 1: Train a critic LLM by distilling GPT-4 labeling results of reflection tokens, reaching over 90% consistency with human annotation.

Stage 2: Expand generator vocabulary with all reflection tokens, inject labeled reflection signals into instruction data offline, and jointly predict text content and reflection tokens under standard next-token loss. Different from RLHF, all critique signals are pre-computed offline without real-time reward models, largely lowering training cost.

3.3.3. Inference algorithm & controllable decoding

SELF-RAG inference workflow: Input query and historical text → predict retrieve token → trigger retrieval if needed → generate multiple candidate segments paired with ISREL/ISSUP/ISUSE scores → rank candidates via weighted sum of critique scores → output optimal segment.

Segment-level weighted beam search is supported at the inference stage: users adjust the weight of the ISSUP token to prioritize fully-evidenced factual answers or raise the ISUSE weight for fluent creative writing. An adaptive retrieval threshold can be set to control retrieval frequency for

DOI: 10.54254/2755-2721/2026.GL35629

different tasks. Experimental results on open-domain QA and fact verification show SELF-RAG outperforms ChatGPT and Llama2 RAG on factual accuracy and citation precision, with only a 2% unsupported output ratio compared to 15%-20% of baseline chat LLMs.

3.4. Knowledge graph enhanced anti-hallucination technology (KRAGEN)

KRAGEN integrates knowledge graph (KG), RAG and Graph-of-Thought (GoT) prompting for complex biomedical reasoning tasks, solving the problem that plain text RAG cannot model entity-relation logical constraints, which reduces reasoning hallucinations in multi-hop professional questions.

3.4.1. KG vectorization based on weaviate

KRAGEN converts Neo4j knowledge graph dumps into natural language statements, then embeds all entities and relations into vector space via embedding models, and stores vectors in the Weaviate vector database supporting hybrid keyword-semantic search. Users deploy the whole pipeline via one Docker Compose command, compatible with structured and unstructured domain data.

3.4.2. Graph-of-Thought reasoning mechanism

GoT decomposes complex domain questions into interconnected subproblems (graph vertices), and regards reasoning dependencies as edges between subproblems. Each subproblem calls the internal RAG module to retrieve domain KG knowledge and generates sub-answers; sub-results are delivered to subsequent reasoning nodes until the final comprehensive solution is synthesized. The GoT visualization GUI displays all reasoning vertices, retrieved knowledge and intermediate outputs for human inspection of logical flaws and hallucinations.

3.4.3. Biomedical dataset experiments

The Alzheimer's disease knowledge graph AlzKB is used for validation, containing genes, disease and drug entities with relational links. Compared with standard flat prompting RAG, KRAGEN lifts 1-hop reasoning accuracy from 56.6% to 70.4%, and 2-hop reasoning accuracy from 53.1% to 71.8%. True/false biomedical QA accuracy also rises from 68.6% to 80.3%, while complex multi-hop questions only gain slight improvement. KRAGEN surpasses BioGPT and general GPT baselines on all medical reasoning tasks.

4. Comprehensive comparison of two RAG evaluation benchmarks

To quantify hallucination severity and overall RAG reliability, two authoritative benchmarks, CRAG and RAGEval, are designed for distinct evaluation purposes, with obvious differences in dataset construction, metrics and applicable scenarios.

4.1. CRAG benchmark

CRAG contains 4,409 English QA pairs covering five domains (Finance, Sports, Music, Movie, Open) and eight query categories including multi-hop, aggregation and false-premise questions. It divides facts into four temporal groups: second-level real-time content, daily fast-changing

DOI: 10.54254/2755-2721/2026.GL35629

information, yearly slow-updating facts and static knowledge, supported by 220K HTML web pages and a simulated knowledge graph with 2.6M entities.

CRAG uses a four-grade manual scoring rule: Perfect (1 point, no hallucinations), Acceptable (0.5 point, minor errors), Missing (0 point, no valid answer), Incorrect (-1 point, conflicting hallucinations), imposing heavy punishment on fabricated content. Automatic evaluation via dual ChatGPT and Llama 3 judges achieves 94.7%–98.9% F1 [9]. Baseline tests show GPT-4 Turbo only reaches 33.5% factual accuracy without RAG; simple fixed Top-K RAG raises it to 43.6%, while mainstream commercial RAG systems have merely 63% fully correct answers and a 16%–25% hallucination ratio. All models drop sharply in accuracy on tough cases with real-time facts, long-tail entities and multi-hop reasoning. CRAG supports three evaluation tasks: retrieval summarization, KG-web hybrid retrieval and noisy end-to-end RAG ranking, suitable for testing real-time dynamic knowledge and long-tail hallucinations of commercial RAG products.

4.2. RAGEval & dragonball benchmark

Different from manually built CRAG, RAGEval is an automatic benchmark generation framework following the pipeline: Schema → Configuration → Document → Q&A → Reference → Fact Key Points. It produces the DragonBall bilingual benchmark with 6,711 QA samples across finance, law, and medical domains, and seven question types, including unanswerable queries. Three core semantic metrics are proposed instead of traditional lexical ROUGE/BLEU: Completeness: Ratio of ground-truth key points covered by the generated answer; Hallucination: Ratio of key points contradicted by output; Irrelevance: Ratio of unrelated key points without coverage or contradiction [5].

The three metrics focus on factual semantics rather than text overlap. Experiments prove Baichuan-7B achieves high ROUGE-L (38.30%) but low Completeness (60.25%), while GPT-4o has low lexical scores yet top Completeness (79.13%). Human annotation verifies LLM automatic scoring has a Pearson correlation over 0.75 with human labels, and an absolute error of only 1.67. Ablation on chunk size and Top-K retrieval shows optimal parameters differ by question types (FQ/MRQ/NCQ require distinct chunk-TopK combinations), without a universal configuration for all tasks. RAGEval fits custom vertical domain evaluation and automatic dataset construction for academic ablation experiments.

4.3. Benchmark comparison summary

CRAG relies on human-annotated real-world dynamic facts, targeting industrial commercial RAG product reliability testing, with a clear heavy penalty on hallucinated answers; RAGEval automatically generates bilingual multi-domain evaluation sets via schema rules, provides three semantic fact-based metrics and is more suitable for academic algorithm ablation and vertical domain customized benchmark building.

5. General industrial architecture of stacked anti-hallucination technologies

Integrating the four technical branches, we summarize a universal layered anti-hallucination RAG architecture as follows:

Retrieval Layer: Blended multi-index hybrid retrieval plus retrieval reordering filters hard negatives and boosts retrieval precision;

DOI: 10.54254/2755-2721/2026.GL35629

Training Layer: Asynchronous full end-to-end RAG joint training with statement reconstruction auxiliary loss enables vertical domain adaptation;

Generation Layer: SELF-RAG self-reflection token mechanism implements dynamic retrieval triggering and segment-level factual self-verification;

Professional Reasoning Layer (Medical/Finance/Legal): KRAGEN KG+GoT module models entity-relation logical constraints to eliminate multi-hop reasoning hallucinations.

The evaluation module can adopt CRAG for real-time dynamic fact testing or RAGEval for customized domain ablation per business needs. All modules are orthogonal and can be selectively combined based on computing resources and domain traits. Experimental results from ten papers prove this multi-layer stacked pipeline achieves higher factuality and lower hallucination ratios than single optimization schemes.

6. Common deficiencies of current research

Lack of unified lightweight anti-hallucination frameworks: Most joint training and self-reflection architectures consume massive GPU resources; few zero-fine-tune lightweight schemes are explored for edge devices. Absence of universal cross-domain evaluation standards: CRAG and RAGEval adopt inconsistent scoring rules, with no unified factual metrics recognized by academia and industry. Insufficient multimodal RAG research: Current anti-hallucination methods focus on plain text and rarely support image, audio and video retrieval scenarios. Fragmented vertical domain research: Medical and software RAG are fully studied, while agriculture, manufacturing and other industries lack sufficient research. Most studies adopt open-source LLMs and seldom discuss costs and hallucination risks of closed API models; ethical and privacy risks are barely covered in anti-hallucination research: KGs and domain corpora contain sensitive medical, legal and private data prone to leakage during retrieval and generation [10].

7. Conclusion

This paper systematically organizes the RAG hallucination research challenges proposed in the introduction, including dispersed suppression methods, insufficient cross-method comparisons, inconsistent evaluation metrics and fragmented vertical-domain solutions. It generalizes the shared fundamental causes of RAG hallucinations, compares four mainstream mitigation strategies (retrieval denoising, domain-adaptive training, generative self-verification, knowledge graph augmentation), and elaborates each strategy's applicable range and anti-hallucination mechanisms. The paper also distinguishes the functional orientations of CRAG and RAGEval evaluation tools, puts forward a universal industrial deployment framework combining multiple hybrid technologies, and concludes prevalent flaws of existing studies to fulfill its initial research goals.

This survey still has clear limitations: it only performs theoretical summary from published papers without self-designed ablation experiments to validate comparative results, and contains no dedicated analysis for lightweight edge device use cases. Future research will incorporate multimodal RAG literature, supplement quantitative indicator comparisons and lightweight scheme analyses, focusing on three key directions: lightweight fine-tune-free hallucination suppression frameworks to cut deployment computing overhead; unified general quantitative evaluation criteria for RAG hallucinations; multimodal and cross-industry knowledge graph-enhanced RAG systems to make up insufficient scenario coverage, so as to build a reliable full-scene retrieval-augmented large language model ecosystem.

DOI: 10.54254/2755-2721/2026.GL35629

References

[1] Sawarkar, K. (2024). Blended RAG: Improving RAG (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *IEEE, California*.
[2] Siriwardhana, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *TACL, 11*, 1–17.
[3] Asai, A. (2024). SELF-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*. Vienna.
[4] Matsumoto, N. (2024). KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics, 40*(6).
[5] Zhu, K. L. (2025). RAGEval: Scenario specific RAG evaluation dataset generation framework. In *ACL*, Vienna.
[6] Bowen, J. (2025). Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. *International Conference on Learning Representations*, Singapore.
[7] Fan, W. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *KDD*, Barcelona.
[8] Tural, B. (2024). Retrieval-augmented generation (RAG) and LLM integration. In *ISAS*, Istanbul.
[9] Yang, X. (2024). CRAG–Comprehensive RAG benchmark advances in neural information processing systems. In *NeurIPS*, Vancouver.
[10] Arslan, M. (2024). A survey on RAG with LLMs. In *KES*, Seville.