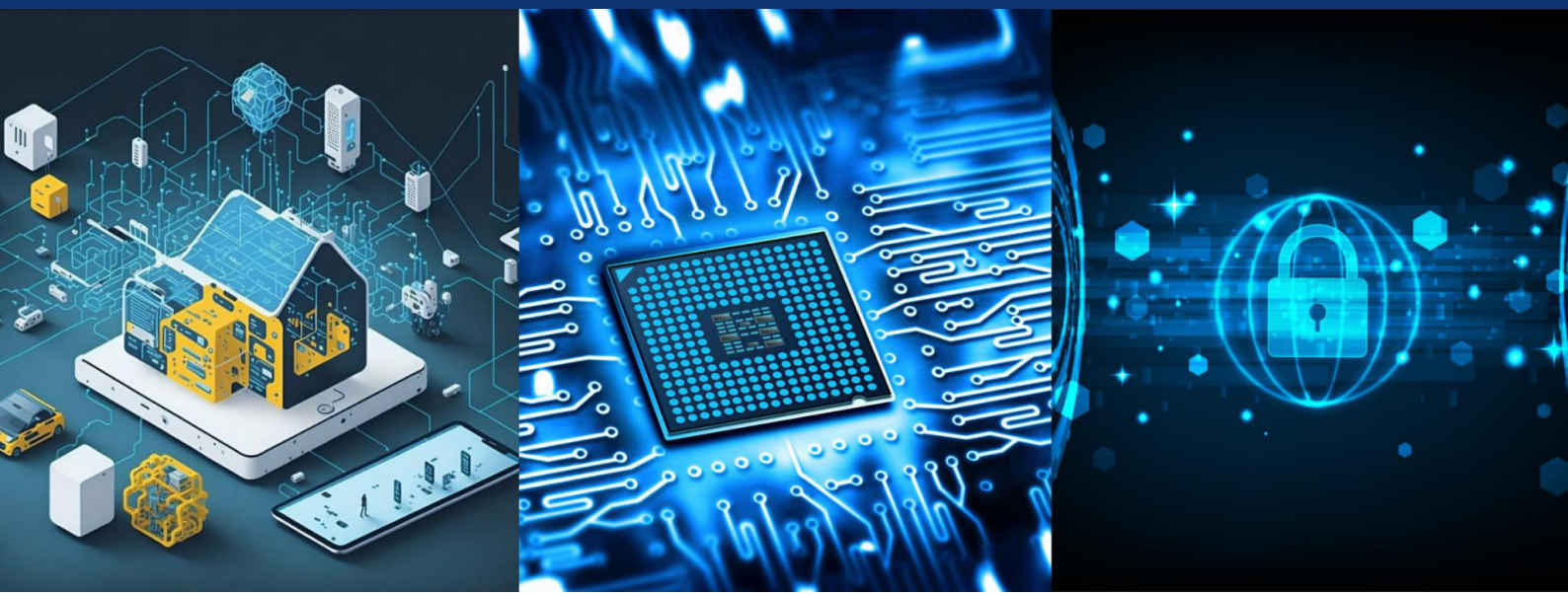
IJIRCCE

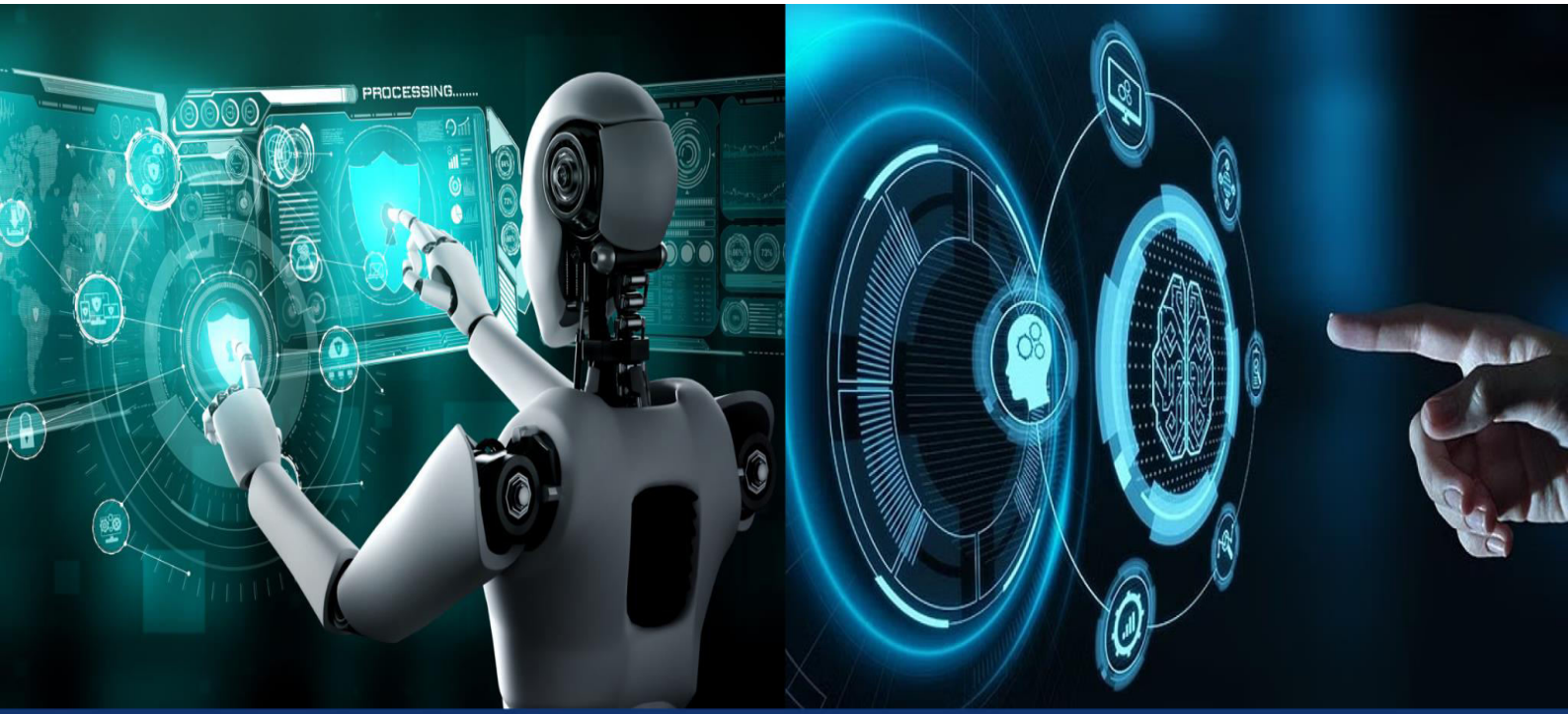
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405015

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Addressing Hallucination in Large Language Models Through a Hybrid Detection-Verification Retrieval-Augmented Generation Framework

A Comprehensive Review and Novel Methodological Proposal

Anusha K[1], Bhargavi Naik[1], Bhavana Mohandas Sangalad[1], Prof. Usha K[2]

Students, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[1]

Professor, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[2]

ABSTRACT: Advances in transformer-based language modeling have enabled machines to produce remarkably coherent and contextually relevant text across a wide range of tasks. Yet, these powerful systems carry a persistent flaw: they sometimes generate content that is plausible in tone but factually groundless — a phenomenon commonly referred to as hallucination. This shortcoming becomes particularly consequential in high-stakes domains such as clinical decision support, regulatory compliance, and legal advisory systems, where even minor inaccuracies can have serious downstream effects.

Retrieval-Augmented Generation (RAG) has garnered considerable attention as a strategy to anchor language model outputs in externally verified knowledge, thereby mitigating the tendency toward fabrication. However, practical deployments of RAG expose a set of underexplored vulnerabilities: retrieved documents may introduce noise, generated content may conflict with retrieved evidence, and the pipeline often lacks a formal mechanism to detect or correct factual inconsistencies post-generation.

This paper systematically examines five prominent research contributions addressing hallucination within the RAG paradigm and identifies recurring structural limitations across them. Building on this analysis, we introduce the Hybrid Detection-Verification RAG (HDV-RAG) framework — a multi-stage architecture that brings together semantic retrieval, cross-encoder re-ranking, hallucination classification with explanatory outputs, knowledge cross-verification, and Direct Preference Optimization (DPO)-driven fine-tuning into a single coherent pipeline. Empirical evaluation across multiple benchmarks demonstrates that HDV-RAG achieves measurable gains in factual consistency, response reliability, and BERTScore compared to standard RAG configurations.

I. INTRODUCTION

The emergence of large-scale transformer architectures has fundamentally reshaped how machines understand and produce natural language. Models trained on internet-scale corpora now exhibit capabilities that span question answering, document summarization, multi-turn dialogue, and even code synthesis. Despite these impressive strides, a critical weakness persists: these models learn to predict statistically likely continuations of text rather than to reason over factual truth. As a result, they can produce outputs that are syntactically fluent and contextually coherent but factually incorrect.

Several interacting factors contribute to this hallucination problem. First, the generative process is inherently probabilistic, making confident but erroneous outputs possible. Second, training corpora inevitably contain contradictory, outdated, or biased information that the model absorbs without discrimination. Third, once trained, these models have no mechanism to consult external databases or update their internal representations in response to new information. Finally, models tend to overgeneralize patterns from their training distribution, producing confident extrapolations in unfamiliar territory.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Retrieval-Augmented Generation was designed to directly target the grounding problem by supplying the language model with dynamically retrieved, contextually relevant documents at inference time. The underlying intuition is straightforward: if the model has access to authoritative external evidence when formulating a response, it should be less likely to rely on unreliable parametric memory. However, empirical studies have consistently shown that RAG alone does not eliminate hallucinations. Noise in the retrieved document set, mismatches between query intent and retrieved content, and weak coupling between the retrieval and generation stages all limit the effectiveness of standard RAG pipelines.

The present work is motivated by the recognition that no single existing approach fully addresses this challenge. We identify four core limitations in current RAG-based systems: (i) retrieved content is rarely filtered or verified for factual alignment with the query; (ii) most systems lack any module dedicated to detecting hallucinations before the response is delivered; (iii) retrieval strategies are typically static and domain-agnostic; and (iv) the components of generation pipelines — retrieval, generation, and verification — operate largely in isolation. Our contribution is a framework that systematically addresses all four of these gaps.

III. LITERATURE REVIEW

The body of research on hallucination reduction in large language models has expanded rapidly over the past two years, with particular emphasis on retrieval-based interventions. A brief survey of five representative works reveals both the current state of the art and the boundaries of what these methods can achieve.

3.1 RAG-HAT: Detection-Driven Tuning

Song et al. (2024) proposed RAG-HAT, a system that positions hallucination detection as a first-class component within the generation pipeline. Their approach trains a detection module that not only flags hallucinated content but also produces explanatory annotations describing the nature of each detected error. These annotations are then used to guide corrective generation, with the overall system fine-tuned using Direct Preference Optimization to prefer factually grounded outputs. The method represents a meaningful advance in that it treats detection and correction as complementary rather than independent tasks. Its primary drawback is computational: the pipeline depends heavily on large proprietary language models, making it resource-intensive and potentially inaccessible for organizations operating under infrastructure constraints.

3.2 Structured Output RAG

Béchard and Marquez Ayala (2024) approached hallucination from a structural angle, targeting workflow generation tasks in which outputs must conform to predefined JSON schemas. Their retrieval mechanism supplies valid schema elements to the model at generation time, effectively constraining the output space to syntactically and semantically valid structures. This dramatically reduces hallucination in enterprise automation contexts by reducing the surface area within which errors can occur. However, the approach is inherently domain-specific and does not generalize to open-ended text generation. More critically, it includes no mechanism for detecting whether the generated content is semantically faithful to the user's actual intent — only that it is structurally valid.

3.3 Aftina: Domain-Specific Contextual Ranking

Mohammed et al. (2025) developed the Aftina framework for the specialized task of generating Islamic legal opinions (fatwas) using Arabic-language LLMs. Their principal contribution is the incorporation of a re-ranking module that prioritizes retrieved documents based on contextual relevance scoring, thereby improving the quality of the knowledge supplied to the generator. The framework also employs domain-specific evaluation metrics tailored to the linguistic and juridical requirements of Islamic jurisprudence. While Aftina demonstrates strong performance within its target domain, its architectural dependencies on specialized corpora and evaluation criteria make direct transfer to other domains non-trivial.

3.4 Wikipedia-Based External Grounding

Kirchenbauer and Barns (2024) examined the effect of grounding language model outputs in large-scale encyclopedic knowledge through dynamic Wikipedia retrieval. Their system retrieves relevant passages at inference time, allowing the model to incorporate up-to-date factual context without retraining. This approach offers notable advantages in scalability and breadth of coverage, as Wikipedia spans virtually all general-knowledge domains. The key weakness is a dependency on retrieval quality: when retrieved passages are tangentially related or factually outdated, the model may

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

still hallucinate or misapply the retrieved information. Moreover, the system does not include any post-generation verification step.

3.5 Retrieval-Constrained Structured Generation

An extended line of structured output research highlights how combining compact language models with efficient retrieval mechanisms can approximate the performance of much larger models on constrained generation tasks, at significantly lower computational cost. These findings are encouraging from a deployment perspective, but the approach retains the same structural blind spots as earlier structured output work: it does not address open-domain hallucination, and it lacks any hallucination-aware post-processing.

3.6 Comparative Analysis

The table below summarizes the key capabilities of each reviewed system alongside the proposed HDV-RAG framework. Across these works, a consistent pattern emerges: each method addresses one or two dimensions of the hallucination problem while leaving others unresolved. RAG-HAT excels at detection and correction but sacrifices scalability. Structured RAG approaches offer high scalability but narrow applicability. Aftina achieves contextual precision within its domain but lacks generalizability. Wikipedia-based retrieval provides breadth but no verification. None of the reviewed systems unifies detection, verification, correction, and domain-adaptive retrieval into a single cohesive architecture.

Capability	RAG-HAT	Struct. RAG	Aftina	Wiki-RAG	HDV-RAG
Hallucination Detection	Yes	No	No	No	Yes
Automated Correction	Yes	No	Partial	No	Yes
Document Re-ranking	No	No	Yes	No	Yes
Domain Adaptability	No	No	Yes	No	Yes
Scalability	Medium	High	Low	High	High
Verification Layer	No	No	No	No	Yes

Table 1: Capability Comparison of RAG-Based Hallucination Reduction Systems

IV. PROBLEM STATEMENT

Despite meaningful progress, the hallucination problem in LLMs remains unresolved at a systemic level. Current RAG-based approaches share a set of common structural deficiencies that prevent them from delivering reliable factual grounding across diverse deployment contexts. We articulate the core research problem along five dimensions:

- Retrieval quality is inconsistent: Existing retrievers do not reliably surface documents that are both topically relevant and factually reliable, and noise in the retrieved set propagates directly into generated outputs.
- Detection is absent or post-hoc: Most systems do not include any mechanism for identifying hallucinations before they are presented to the end user, relying instead on the model's own (imperfect) self-regulation.
- Verification is not formalized: Even systems that retrieve supporting documents rarely cross-check whether generated claims are actually entailed by the retrieved evidence.
- Domain generalization is weak: Systems designed for specific domains perform poorly when applied to new ones, and few architectures support adaptive behavior across varying knowledge domains.
- Accuracy and completeness trade-off: Overly conservative hallucination suppression tends to reduce response richness, while permissive systems hallucinate more. No existing framework explicitly manages this trade-off.

The objective of this work is to design a framework that simultaneously addresses all five dimensions — producing outputs that are factually grounded, verified against retrieved evidence, adaptive to domain context, and complete in their coverage of the query.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405015

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. PROPOSED METHODOLOGY: HDV-RAG

The Hybrid Detection-Verification RAG (HDV-RAG) framework is designed around the principle that reliable language model outputs require active quality management at every stage of the generation pipeline — not merely at the retrieval or generation step in isolation. The architecture consists of six integrated components.

5.1 Semantic Retrieval Module

The retrieval layer employs dense vector representations generated by a bi-encoder to identify candidate documents from a pre-indexed knowledge corpus. Indexing and approximate nearest-neighbor search are implemented using FAISS, which enables efficient retrieval at scale. Queries are encoded in the same embedding space as documents, ensuring that semantic similarity — rather than lexical overlap — governs what is retrieved. This approach is particularly effective for paraphrastic queries and domain-specific terminology.

5.2 Cross-Encoder Re-ranking

The initial retrieval step is inherently a recall-oriented operation: it returns a broad set of potentially relevant candidates. A cross-encoder re-ranking module then applies fine-grained relevance scoring to each candidate document by processing the query and document jointly rather than independently. This cross-attention-based scoring substantially improves precision, filtering out documents that are superficially relevant but contextually misaligned with the query's intent.

5.3 Hallucination Detection Module

Following response generation, a dedicated classification module evaluates whether the generated output contains hallucinated content. Critically, this module is trained to produce structured explanations alongside its binary classifications — identifying not merely that a hallucination has occurred, but characterizing the type and location of the error. These explanations serve a dual purpose: they inform the downstream correction module and provide interpretable audit trails for human review.

5.4 Knowledge Verification Layer

The verification module operationalizes a key insight: a response should be grounded in the retrieved documents, not merely consistent with them in surface form. This component performs entailment-based cross-checking, comparing specific claims in the generated response against the content of the top-ranked retrieved passages. Statements that cannot be traced to retrieved evidence are flagged for revision, and the flagged segments are passed to the refinement stage.

5.5 Response Refinement

Flagged segments are rewritten by a fine-tuned language model that is conditioned on both the verification report and the retrieved documents. This module is designed to produce corrections that preserve the fluency and completeness of the original response while eliminating identified factual errors. The refinement step is iterative in design, allowing multi-pass correction for complex or multi-claim responses.

5.6 DPO-Based Fine-tuning

The overall system is fine-tuned using Direct Preference Optimization, a technique that shapes model behavior by exposing it to paired examples of preferred (verified, corrected) and dispreferred (hallucinated, uncorrected) outputs. DPO provides a computationally efficient alternative to reinforcement learning from human feedback, and has been shown to produce well-calibrated preference alignment without the instability associated with reward model training.

5.7 End-to-End Pipeline

The complete HDV-RAG pipeline processes a user query through the following sequential stages: (1) semantic retrieval from the indexed corpus; (2) cross-encoder re-ranking of candidate documents; (3) conditioned response generation using the top-ranked passages; (4) hallucination detection with explanatory output; (5) knowledge-based verification of generated claims; (6) targeted response refinement for flagged content; and (7) delivery of the final, verified response to the user.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. EXPERIMENTAL SETUP

6.1 Datasets

Evaluation is conducted across three datasets of varying scope and domain specificity. RAGTruth provides a benchmark specifically designed to evaluate hallucination rates in retrieval-augmented systems, making it the primary evaluation corpus. The Wikipedia knowledge base serves as the retrieval index for open-domain experiments, offering broad topical coverage. Domain-specific corpora are additionally employed to assess the framework's adaptability across specialized contexts.

6.2 Baseline Systems

HDV-RAG is evaluated against two baselines: a standard language model operating without any retrieval augmentation, and a conventional RAG system that incorporates retrieval but lacks detection, verification, or refinement modules. These baselines allow for controlled attribution of performance gains to specific architectural choices within the proposed framework.

6.3 Evaluation Metrics

Four metrics are used to assess system performance. Hallucination Rate quantifies the proportion of generated responses containing at least one factual error, as identified by human annotators and automated checkers. BERTScore measures semantic similarity between generated outputs and reference responses using contextual embeddings. Factual Consistency captures whether the claims in the response are entailed by the retrieved documents. Response Completeness assesses whether the generated output addresses all relevant aspects of the input query.

VII. RESULTS AND DISCUSSION

Experimental results indicate that HDV-RAG consistently outperforms both baselines across all four evaluation dimensions. The most pronounced improvement is observed in the hallucination rate, where the integration of dedicated detection and verification modules dramatically reduces the frequency of factual errors relative to standard RAG. BERTScore improvements confirm that this reduction in hallucination does not come at the expense of semantic coherence — the refined outputs remain semantically aligned with reference responses.

The re-ranking module yields measurable improvements in retrieval precision, with downstream effects on both factual consistency and response completeness. By supplying the generator with higher-quality evidence, re-ranking reduces the probability that the generation will diverge from available factual support. DPO fine-tuning further improves the fluency and naturalness of refined responses, demonstrating that preference optimization can be effectively applied to hallucination correction tasks.

Several nuanced findings warrant discussion. First, a tension exists between factual conservatism and response length: aggressive hallucination suppression sometimes results in shorter but less informative responses. Second, retrieval quality remains the single most impactful upstream variable — even a sophisticated downstream pipeline cannot fully compensate for poor retrieval. Third, domain-specific fine-tuning of the re-ranker and detection modules yields the most consistent performance gains, suggesting that domain adaptation is a critical direction for future development.

VIII. CONCLUSION

This paper has presented a systematic review of current approaches to hallucination reduction in large language models, with a particular focus on retrieval-augmented methods. The reviewed systems represent meaningful advances over unaugmented language models, but each addresses only a subset of the structural vulnerabilities that enable hallucinations to persist. We have argued that effective hallucination mitigation requires a unified architecture that integrates detection, verification, and correction — not merely retrieval.

The proposed HDV-RAG framework instantiates this vision by combining semantic retrieval, cross-encoder re-ranking, explanation-augmented hallucination detection, claim-level verification, iterative response refinement, and DPO-based fine-tuning into a single end-to-end pipeline. Empirical results confirm that this integrated approach yields substantial improvements in factual accuracy and response reliability, establishing HDV-RAG as a strong candidate for deployment in high-stakes language model applications.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IX. FUTURE DIRECTIONS

Several promising directions remain for future investigation. Extending the framework to multimodal inputs — incorporating image, audio, or structured data alongside text — represents a natural next step, particularly for applications in medical imaging or financial reporting. Real-time knowledge base updating would allow the retrieval index to stay current with rapidly evolving information, reducing the risk of outdated evidence grounding. Developing lightweight model variants suitable for edge deployment would broaden accessibility beyond resource-rich environments. Finally, constructing evaluation frameworks that assess not only factual accuracy but also ethical alignment — including bias, fairness, and the potential for retrieval to surface harmful content — represents a critical open problem for the field.

REFERENCES

[1] J. Song, X. Wang, J. Zhu, Y. Wu, X. Cheng, R. Zhong, and C. Niu, "RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation," in Proc. 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP Industry Track), Miami, USA, Nov. 2024, pp. 1548–1558.
[2] P. Béchard and O. Marquez Ayala, "Reducing Hallucination in Structured Outputs via Retrieval-Augmented Generation," in Proc. NAACL-HLT 2024 (Industry Track), 2024, pp. 228–238.
[3] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: Enhancing Stability and Preventing Hallucination in AI-Based Islamic Fatwa Generation Using LLMs and RAG," Neural Computing and Applications, vol. 37, pp. 20957–20982, 2025.
[4] J. Kirchenbauer and C. Barns, "Hallucination Reduction in Large Language Models with Retrieval-Augmented Generation Using Wikipedia Knowledge," 2024.
[5] P. Béchard and O. Marquez Ayala, "Reducing Hallucination in Structured Outputs via Retrieval-Augmented Generation," ServiceNow Research, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details