Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202511

Research on Hallucination Mitigation in Large Language Models

Yihan Wang*

College of Sciences, Nanjing Agricultural University, Nanjing, 210000, China

Corresponding author: Yihan Wang*

Abstract

The proliferation of large language models has brought the issue of hallucinations, which are outputs that deviate from inputs or fabricate information, to the forefront of concerns regarding reliability and deployment safety. This paper provides a systematic review of existing mitigation strategies and evaluation frameworks for hallucinations. Among mitigation approaches, supervised fine-tuning and reinforcement learning from human feedback have demonstrated moderate success; however, their heavy reliance on extensive, high-quality human annotations renders them costly and difficult to scale. On the evaluation side, this study examines several multimodal benchmarks, including POPE, MMHal-Bench, MM-Vet, and MMBench, which enable comprehensive assessment of model outputs through a blend of automated metrics and human judgment across multiple dimensions. Overall, substantial progress has been achieved in hallucination evaluation frameworks, yet scalable and cost-effective mitigation techniques remain elusive, particularly for hallucinations in complex multimodal settings, where further breakthroughs are urgently needed.

Keywords

large language models, hallucination mitigation, multimodality

1. Introduction

The rapid advancement of generative artificial intelligence in recent years has positioned large language models as powerful tools with remarkable intelligence and generalization capabilities, unlocking substantial potential across diverse industries (Li et al., 2025). Despite these strengths, such models frequently exhibit “hallucinations” during generation and reasoning—producing outputs that diverge from the input or are entirely fabricated. In practice, this phenomenon undermines system reliability and user trust while posing considerable risks in real-world applications. From an academic perspective, in-depth investigation of multimodal hallucinations and the development of effective mitigation strategies hold significant value. Moreover, enhancing the factual accuracy of model outputs and ensuring safety in deployment render this line of inquiry critically important. Recent work has introduced a hallucination detection approach tailored to retrieval-augmented private question-answering models, which integrates uncertainty-based metrics with automated evaluation by the language model itself, leveraging both intrinsic generation patterns and the model’s self-assessment capabilities (Zhang, 2025). The present study evaluates this method against three distinct baselines, highlighting an imbalance in the literature: while evaluation methodologies for hallucinations in large language models have progressed toward systematic frameworks, they continue to face

substantial challenges. Collectively, these observations underscore the necessity of further exploration in this domain.

2. Research Methods and Data

2.1 Mainstream Research Methods

Current efforts to mitigate hallucinations in generative models predominantly rely on supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Zheng, 2025). SFT leverages extensive, high-quality human-annotated data to refine the model, thereby enhancing its grasp of real-world semantics and commonsense knowledge, which in turn reduces the generation of erroneous information. In contrast, RLHF incorporates human judgments into the optimization loop by iteratively aligning outputs with human preferences, yielding further improvements in factual accuracy and plausibility. Although these approaches have achieved measurable success in curbing hallucinations, their practical deployment reveals notable limitations. Both methods depend critically on high-quality annotations—a resource that is not only costly to acquire but also challenging to scale. For multimodal tasks, the annotation process is particularly complex and susceptible to subjective bias, complicating the creation of robust training datasets. Moreover, RLHF necessitates the development of a reward model and sustained interaction with human feedback, rendering the training pipeline intricate and hindering rapid adoption. Consequently, devising strategies that minimize reliance on human annotations while streamlining the training process—without compromising hallucination control—emerges as a pressing challenge in the field.

2.2 Datasets and Evaluation Metrics

POPE serves as a widely adopted evaluation protocol for hallucination mitigation (Li et al.). Tailored to large-scale vision-language models (LVLMs), its core design requires the model to respond only with “yes” or “no” to simple binary questions, thereby probing for hallucinatory outputs. For instance, questions are framed as “Does the image contain a car?” and constructed by introducing objects that either do or do not appear in the image, yielding a binary classification task. The POPE dataset comprises 3,000 such questions derived from 500 images. Model responses are treated as binary predictions, enabling the computation of accuracy, precision, recall, and F1 score. Higher values across these metrics indicate reduced hallucination propensity. Three sampling strategies, namely random, popularity-based, and adversarial, are incorporated to generate negative examples, differing primarily in how absent objects are selected.Random sampling draws from objects never present in the image; popularity sampling selects high-frequency objects absent from the specific image; and adversarial sampling targets entities that frequently co-occur in typical contexts but are missing in the given instance. This framework provides a robust mechanism for quantifying hallucinations in multimodal model outputs, a principle underpinning the MMHal-Bench benchmark (Sun et al., 2024).

The MMHal-Bench dataset consists of 8 categories, each containing 12 image groups, yielding a total of 96 groups. During evaluation, the image, model response, and human-annotated ground truth are scored by GPT-4 on a 0–5 scale; responses below 3 are classified as hallucinatory.

MM-Vet is designed to assess the integrated capabilities of multimodal large language models on complex tasks. It encompasses six core competencies: recognition, knowledge comprehension, optical character recognition (OCR), spatial awareness, language generation, and mathematical reasoning. Using GPT-4, model outputs are automatically compared against human-provided reference answers across a dataset comprising 200 images and 218 questions, with performance quantified through scoring.

MMBench, centered on perception and reasoning, further decomposes these into 6 second-level dimensions and 20 third-level subdimensions, spanning image style, scene understanding, emotion recognition, attribute judgment, logical reasoning, and beyond, thus establishing a comprehensive capability evaluation framework. In total, the benchmark includes 3,217 samples, split in a 4:6 ratio into development and test sets. The development set, along with its answers, is publicly available, whereas only the questions are released for the test set. In subsequent experiments, 500 samples are randomly drawn from the development set for model performance evaluation.

To systematically examine the effectiveness of hallucination reduction and the preservation of original

multimodal proficiency, this study employs two evaluation paradigms: (1) Hallucination suppression metrics: precision and F1 score are measured on the POPE dataset under all three sampling conditions, supplemented by MMHal-Bench scores and hallucination frequency; (2) Multimodal general capability metrics: aggregate scores across the six dimensions of MM-Vet (recognition, knowledge comprehension, OCR, spatial relation judgment, language generation, and logical reasoning), alongside consolidated perception and reasoning results from MMBench.

3. Experimental Results and Analysis

To systematically evaluate the performance of the proposed method, three baseline models are selected as comparisons in this study, with experimental conditions configured accordingly. First, LLaVA-1.5 serves as the unoptimized vanilla benchmark, reflecting the initial capabilities of the base architecture. Second, HA-DPO implements direct preference optimization using 16,000 preference pairs automatically annotated by GPT-4, representing a mainstream approach to multimodal hallucination suppression. Third, HDPO conducts targeted optimization for specific hallucination categories. It systematically constructs a training dataset encompassing three major problem types, namely visual distractor interference, long-text comprehension bias, and multimodal semantic inconsistency, with a data scale of 20,000 samples, enabling precise refinement of recognized failure modes and enhancing semantic understanding in complex scenarios.

The core value of this strategy lies in its ability to deliver targeted improvements for identified hallucination patterns, thereby effectively strengthening the model's semantic understanding in complex contexts. For hallucination evaluation, this study compares different baseline methods through experimental results on two primary metrics: POPE and MMHal. To assess the model's overall multimodal capabilities, additional tests are conducted on the MMBench and MM-Vet benchmarks. Specifically, Table 1 reports accuracy, F1 scores, and yes-response rates across three sampling conditions in POPE, while Table 2 presents performance in terms of MMHal scores and their corresponding hallucination rates.

Table 1: Accuracy and F1 Score on the POPE Hallucination Benchmark

POPE Sampling	Method	Accuracy	F1 Score
Random	LLaVA-1.5	86	85.71
	HA-DPO	88.7	80.26
	HDPO	89.6	89.67
Popularity	LLaVA-1.5	76.67	78.79
	HA-DPO	83.36	84.33
	HDPO	86.2	86.79
Adversarial	LLaVA-1.5	73.33	76.05
	HA-DPO	74.5	78.64
	HDPO	80.76	82.54

Table 2: Mean Score and Hallucination Rate on the MMHal-Bench Hallucination Benchmark

Method	Mean Score	Hallucination Rate
LLaVA-1.5	1.84	0.64
HA-DPO	2.09	0.58
HDPO	2.1	0.55

Tables 1 and 2 present the performance of each method on hallucination benchmarks. The AID-optimized LLaVA-1.5 consistently outperforms the baseline model across all configurations. In the most challenging POPE adversarial setting, it surpasses HA-DPO yet falls slightly behind HDPO, aligning with expectations. HA-DPO's reliance on GPT-4 for high-quality annotations may account for its slightly weaker performance, potentially constrained by the learning capacity of LLaVA-1.5-7B's vision module; scaling to larger models could yield further improvements, though GPT-4's prohibitive cost limits widespread adoption. HDPO leverages distractor prompts to capture hallucination patterns, showing certain advantages in complex scenarios, but the injected noise diminishes the discriminability of preference learning. AID enhances reliability by explicitly delineating objects to anchor model attention to image regions, mitigating language prior interference and improving attribute and relation descriptions, thereby corroborating the efficacy of heightened visual focus for generating reliable descriptions.

4. Current State of the Field

Research on hallucinations in large language models reveals a marked imbalance in that evaluation methodologies have advanced far more rapidly than mitigation strategies. Evaluation frameworks have matured into systematic, hierarchical, and multidimensional systems that integrate automated metrics with human oversight. In contrast, mitigation approaches remain heavily reliant on human intervention and have yet to overcome scalability and cost barriers, challenges that are particularly acute in multimodal and complex reasoning contexts.

In the evaluation domain, concerted efforts have yielded a well-structured benchmark ecosystem capable of quantifying hallucinations at multiple granularities. Fine-grained assessment encompasses sentence-level credibility scoring, factuality verification, and entity-specific truthfulness checks. Holistic output evaluation focuses on coherence, logical consistency, and factual fidelity across entire generations. Task-specific multidimensional profiling examines hallucination patterns in knowledge-intensive question answering, logical reasoning, dialogue generation, and related domains. Established benchmarks such as TruthfulQA, HaluEval, and FACTOR (Lin, 2025) enhance reliability and reproducibility by combining automated signals, such as natural language inference-based fidelity scores and retrieval-augmented fact checking, with targeted human annotation.

In the realm of hallucination mitigation, despite the introduction of diverse techniques, the field remains predominantly anchored in labor-intensive paradigms and continues to grapple with scalability constraints. Supervised fine-tuning (SFT) refines models using high-quality, low-hallucination corpora, whereas reinforcement learning from human feedback (RLHF) aligns generation policies with preference signals. Both approaches demonstrably reduce hallucination rates; however, their dependence on extensive, precise human annotations incurs prohibitive costs and procedural complexity, rendering them ill-suited for rapid iteration or large-scale deployment.

In the pursuit of low-cost alternatives, methods such as leveraging the model's self-supervised capabilities for consistency verification, incorporating retrieval-augmented generation to reduce factual errors, and constructing synthetic data for hallucination mitigation training have proven effective to some extent in specific scenarios. However, their generalization ability remains limited overall, with stability particularly compromised in the face of complex reasoning or dynamic knowledge updates.

With respect to the distinct challenges in multimodal settings, vision-language models must simultaneously address issues including modality alignment errors, absence of cross-modal commonsense knowledge, and contextual understanding biases. Cross-modal hallucinations are more readily induced under adversarial examples or complex reasoning tasks compared to other conditions, yet existing approaches frequently underperform in out-of-distribution generalization. In summary, current research has established a relatively mature framework for hallucination evaluation that clearly guides model improvement. Nevertheless, mitigation techniques continue to depend heavily on high-quality human feedback, rendering scalable and low-cost de-hallucination solutions a critical challenge that awaits breakthrough. To systematically enhance trustworthy generation in a comprehensive and in-depth manner, future investigations should more thoroughly explore lightweight self-correction mechanisms, multimodal alignment techniques, and generation constraint strategies capable of dynamically adapting to knowledge updates.

5. Conclusion

This paper systematically reviews hallucination mitigation techniques and evaluation frameworks for large language models. It critically examines the efficacy and limitations of prevailing approaches, such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). While these methods reduce hallucinations to some extent, their dependence on extensive, high-quality human annotations incurs substantial costs and scalability constraints—challenges that are amplified in complex multimodal scenarios.

On the evaluation front, the study surveys key multimodal benchmarks, including POPE, MMHal-Bench, MM-Vet, and MMBench. These have evolved into systematic, multidimensional frameworks that robustly assess output fidelity and consistency. Through comparative experiments, we demonstrate that optimized methods can suppress hallucinations while preserving—or even enhancing—multimodal comprehension and

reasoning capabilities, underscoring their refinement potential.

Nevertheless, the field of hallucination mitigation remains anchored in human-intensive workflows and has yet to yield scalable, cost-effective, generalizable solutions. Looking ahead, research should prioritize lightweight, adaptive paradigms—such as self-supervised correction mechanisms, improved cross-modal alignment, and dynamic knowledge integration strategies. To ensure trustworthy generation and safe deployment in real-world applications (e.g., conversational agents and writing assistants), future efforts must also bridge evaluation benchmarks with practical use cases, thereby facilitating reliable integration of hallucination mitigation into operational systems.

References

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X. and Wen, J.-R., Published. Evaluating Object Hallucination in Large Vision-Language Models. *The 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Optica Publishing Group, pp. 292-305.

Li, Y. Q., Wang, X. W. and Wang, Q. Y., (2025). Research on multimodal sentiment analysis models for social media based on multimodal large language models. *Information Studies:Theory & Application,* vol. 48**,** no. 11, pp. 188-197.

Lin, Y. W., (2025). *Research on large model hal lucination mitigation methods based on contrastive decoding.* Master's Theses, Guangzhou University.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.-X. and Yang, Y., (2024). Published. Aligning large multimodal models with factually augmented rlhf. *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics (ACL), pp. 13088-13110.

Zhang, J. Q., (2025). *Research on hallucination mitigation in multi-modal large language models.* Master's Theses, University of Electronic Science and Technology of China.

Zheng, H. J., (2025). *From ChatGPT to Deep Seek: A regulatory approach to the risk of illusion in large models* [Online]. Journal of Intelligence. Available: http://kns.cnki.net/kcms/detail/61.1167.g3.20250909.1241.002.html.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access