

Evaluation of Hallucination in Large Language Models: A Literature Survey

Introduction

Hallucination in large language models (LLMs) refers to the generation of plausible but factually incorrect, misleading, or nonsensical content that deviates from the source or reality. As LLMs become increasingly integrated into diverse applications—from medical education and financial analysis to multilingual communication and knowledge-intensive domains—their hallucination tendencies pose significant challenges to reliability, trustworthiness, and safety. Evaluating hallucination effectively is critical for understanding its nature, diagnosing its causes, and developing mitigation strategies. This survey synthesizes recent advances in the evaluation of hallucination in LLMs, drawing on a broad spectrum of methodologies, benchmarks, theoretical analyses, and application-specific studies. The goal is to provide a comprehensive overview of current evaluation techniques, challenges, and emerging directions in the field.

Hallucination Detection Methodologies and Frameworks

Metamorphic Testing and Prompt-Based Detection

Several studies have introduced innovative frameworks for hallucination detection that do not rely solely on external knowledge bases or labeled data. Li et al. (2024) propose **Drowzee**, a metamorphic testing framework that leverages logic programming and a factual knowledge base derived from Wikipedia to detect fact-conflicting hallucinations (FCH). Drowzee automatically generates logic-based test cases with ground truths and uses semantic-aware oracles to assess LLM outputs, revealing high hallucination rates across multiple models and domains. Similarly, Yang et al. (2025) develop **MetaQA**, which employs metamorphic relations and prompt mutation to detect hallucinations without external resources. MetaQA demonstrates superior performance over zero-resource baselines like SelfCheckGPT, showcasing robustness across question types and different LLM architectures.

Liu et al. (2025) introduce **Attention-Guided Self-Reflection (AGSER)**, a zero-shot hallucination detection method that exploits the attention mechanisms within LLMs to split input queries and compute consistency scores, effectively identifying hallucinations with reduced computational cost. Munakata et al. (2024) propose a **multiple-fill-in-the-blank exam** approach, prompting LLMs to repeatedly fill masked blanks, ensuring consistent storylines and enabling fine-grained hallucination quantification at the sentence level.

Retrieval-Augmented and Knowledge-Graph Based Evaluation

Grounding LLM outputs in external knowledge sources has emerged as a powerful approach to hallucination evaluation. Lee and Yu (2025) present **REFIND**,

a retrieval-augmented hallucination detection framework that introduces the Context Sensitivity Ratio (CSR) to measure output sensitivity to retrieved evidence, achieving robust detection across multiple languages. Bashir et al. (2025) analyze the impact of indexing techniques on Retrieval-Augmented Generation (RAG) systems, underscoring the importance of retrieval quality in reducing hallucination in closed-domain question answering.

Graph-based methods further enhance explainability and precision in hallucination evaluation. Sansford et al. (2024) propose **GraphEval**, which leverages knowledge graph structures to pinpoint specific triples causing hallucinations and offers correction mechanisms through **GraphCorrect**. Nishat et al. (2025) introduce **ALIGNed-LLM**, integrating knowledge graph embeddings into LLM latent spaces to improve factual grounding and reduce hallucination, particularly in high-stakes financial applications.

Benchmarking and Dataset Contributions

The development of comprehensive benchmarks is vital for standardized hallucination evaluation. Zhang et al. (2025) introduce **Poly-FEVER**, a large-scale multilingual fact verification dataset covering 11 languages and nearly 78,000 claims, enabling cross-linguistic hallucination analysis and revealing language-specific biases. Liang et al. (2023) present **UHG Eval**, focusing on unconstrained hallucination generation in Chinese LLMs with keyword-level annotations, facilitating realistic and scalable evaluation.

Sun et al. (2024) propose a novel benchmark using **unanswerable math word problems (UMWP)** to evaluate hallucination in question answering, demonstrating that in-context learning and RLHF improve detection of unanswerable queries. Sciré et al. (2024) develop **LLM-Oasis**, the largest end-to-end factuality evaluation resource, combining Wikipedia claims with human-validated falsifications to push the limits of factuality assessment.

In the vision-language domain, Guan et al. (2023) introduce **HALLUSION-BENCH**, diagnosing hallucinations and visual illusions in large vision-language models (LVLMs), revealing significant challenges in grounding visual inputs and mitigating biases.

Theoretical and Semantic Consistency Approaches

Karbasi et al. (2025) provide a theoretical framework demonstrating the inherent difficulty—and in some cases impossibility—of fully automated hallucination detection without expert-labeled negative examples, emphasizing the necessity of human feedback and supervised training.

Rabinovich et al. (2023) approach hallucination evaluation through **semantic consistency** in question-answering, measuring the ability of LLMs to produce semantically equivalent answers to paraphrased questions. Their reference-less

framework predicts answer correctness effectively, offering a promising direction for hallucination detection without external ground truth.

Chen et al. (2023) propose **ReID**, a robust discriminator trained on bilingual QA dialogues enriched with LLM-generated data, which successfully identifies hallucinations across in- and out-of-distribution datasets, contributing to nuanced understanding of hallucination types.

Challenges and Application-Specific Evaluations

Domain-Specific Hallucination Assessment

Multiple studies highlight domain-specific challenges in hallucination evaluation. Silvestri et al. (2024) assess hallucination in a GPT-4-powered chatbot for surgical oral board exams, revealing fabricated clinical findings and inconsistent postoperative explanations, underscoring the risks of hallucinations in medical education and the need for expert oversight.

Baranwal et al. (2024) evaluate hallucination in nursing data standardization, finding GPT-4 significantly reduces hallucinations compared to GPT-3.5, demonstrating the impact of model improvements on domain-specific reliability.

Schmitt et al. (2025) explore ontology-augmented prompting in Alzheimer’s drug repurposing, showing reduced hallucinations and better alignment with expert knowledge, highlighting ontology integration as a mitigation strategy in biomedical applications.

Sharkey and Treleaven (2024) compare BERT and GPT models for financial engineering, noting that while GPT models offer higher predictive power, BERT variants exhibit lower hallucination risk and better interpretability, suggesting model choice considerations based on hallucination profiles.

Hallucination in Multimodal and Vision-Language Models

Cui et al. (2023) analyze hallucination in GPT-4V(ision), identifying bias and interference as key failure modes affecting visual reasoning. Their findings reveal that current mitigation strategies like chain-of-thought reasoning are insufficient, pointing to the need for novel approaches in vision-language hallucination evaluation.

Zhang et al. (2024) focus on **number hallucination** in vision-language models, demonstrating severe inconsistencies in object counting and proposing training methods to improve consistency, contributing to more reliable multimodal outputs.

Zhou et al. (2024) investigate parameter-efficient fine-tuning (PEFT) for multimodal LLMs, showing that certain PEFT methods can reduce hallucination while improving stability and generalization, offering practical insights for multimodal model development.

Hallucination in Interactive and Generative Settings

Jin et al. (2024) study hallucination and logical inconsistencies in LLM-powered text-based games, using automated log analysis to detect hallucination-induced bugs affecting gameplay logic and balance. Their approach enables scalable evaluation without relying on player feedback, pioneering hallucination assessment in interactive AI systems.

Qi et al. (2025) examine mixed-context hallucination in summarization tasks, revealing that LLMs’ intrinsic knowledge biases impair accurate hallucination detection, especially when external context conflicts with internal knowledge, highlighting evaluation complexities.

Xie (2024) investigates the impact of reasoning order on hallucination, introducing a benchmark comparing answer-first versus reasoning-first generation strategies. Reflexive prompting improves factual consistency, offering a practical mitigation technique aligned with evaluation insights.

Mitigation-Oriented Evaluation and Future Directions

Integration with Knowledge Bases and Post-Processing

Li et al. (2025) and Nishat et al. (2025) demonstrate that integrating knowledge graphs into LLM workflows can both evaluate and reduce hallucinations by grounding responses in structured factual data. Li et al.’s **LinkQ** system uses knowledge graph queries during question answering, showing improvements over GPT-4 but also revealing challenges in query construction.

Vladika et al. (2025) explore self-correcting LLM methods in news summarization, using iterative verification and external search engine snippets to refine outputs and reduce hallucinations, highlighting the effectiveness of feedback loops in evaluation and mitigation.

Fang et al. (2025) propose a three-stage correction framework for ASR error correction, combining error pre-detection, chain-of-thought correction, and verification, reducing hallucinations without extra data or fine-tuning, emphasizing evaluation methods that integrate correction and detection.

Surveys and Taxonomies Informing Evaluation

Ji et al. (2022) provide a foundational survey categorizing hallucination types, evaluation metrics, and mitigation strategies, stressing the need for fine-grained and fact-checking-focused metrics.

Kazlaris et al. (2025) and Tonmoy et al. (2024) offer comprehensive taxonomies of hallucination mitigation techniques, highlighting evaluation challenges such as lack of standardized benchmarks and the fragility of retrieval-based approaches. They advocate for hybrid pipelines combining retrieval, generation, and self-reflective reasoning agents, informing future evaluation frameworks.

Abdelghafour et al. (2024) survey mitigation techniques focusing on retrieval augmentation, human feedback, and controlled generation, discussing evaluation hurdles like scalability and data quality dependence.

Theoretical Insights on Hallucination Evaluation

Lee (2023) provides a mathematical analysis linking hallucination to model uncertainty and creativity, suggesting an intrinsic trade-off that complicates complete hallucination elimination without compromising generative diversity. This theoretical perspective informs evaluation by contextualizing hallucination as a fundamental property rather than solely a failure.

Zhang et al. (2025) investigate leveraging hallucinated knowledge for negative reasoning in fake news detection, proposing evaluation methods that exploit hallucination patterns for beneficial downstream tasks, expanding the scope of hallucination evaluation beyond error detection.

Conclusion

Evaluating hallucination in large language models is a multifaceted challenge that spans detection methodologies, benchmark development, domain-specific assessments, and theoretical understanding. Recent advances demonstrate promising directions through metamorphic testing, retrieval-augmented frameworks, knowledge graph integration, and semantic consistency analysis. However, fundamental difficulties remain, including the intrinsic nature of hallucination, variability across languages and domains, and the complexity of mixed-context scenarios. Effective evaluation increasingly relies on combining automated metrics with human expertise and leveraging external knowledge sources. Future research must focus on standardized benchmarks, interpretable evaluation frameworks, and adaptive mitigation strategies that balance hallucination reduction with creativity and performance. This comprehensive survey underscores the critical role of rigorous hallucination evaluation in advancing trustworthy and reliable large language models.

References

- Abdelghafour, M., Mabrouk, M., & Taha, Z. (2024). Hallucination Mitigation Techniques in Large Language Models. *International Journal of Intelligent Computing and Information Sciences*. Unknown URL.
- Baranwal, A., Semenov, A., Salgado, P., Priola, K., Yao, Y., Keenan, G., & Macieira, T. (2024). Leveraging Generative Pre-Trained Transformer Models for Standardizing Nursing Data. *IEEE International Conference on Healthcare Informatics*. Unknown URL.
- Bashir, B., Ibrahim, I., Rabiou, A., & Abdullahi, S. (2025). Analyzing the Impact of Various Indexing Techniques on Retrieval-Augmented Generation (RAG)

Performance in Closed-Domain Question Answering. *International Journal of Science for Global Sustainability*. Unknown URL.

Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., & Xiao, Y. (2023). Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. *International Conference on Information and Knowledge Management*. https://ink.library.smu.edu.sg/context/sis_research/article/9467/viewcontent/3583780.3614905_pv.pdf

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., & Yao, H. (2023). Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges. *arXiv.org*. Unknown URL.

Drowzee: Metamorphic Testing for Fact-Conflicting Hallucination Detection in Large Language Models (Li et al., 2024). Unknown Venue. <https://doi.org/10.1145/3689776>

Fang, Y., Cheng, B., Peng, J., Li, X., Xi, Y., Zhang, C., & Zhong, G. (2025). Fewer Hallucinations, More Verification: A Three-Stage LLM-Based Framework for ASR Error Correction. *arXiv.org*. Unknown URL.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2023). Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *Computer Vision and Pattern Recognition*. <https://arxiv.org/pdf/2310.14566>

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., & Fung, P. (2022). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*. <https://arxiv.org/pdf/2202.03629>

Jin, C., Rao, S., Peng, X., Botchway, P., Quaye, J., Brockett, C., & Dolan, W. (2024). Automatic Bug Detection in LLM-Powered Text-Based Games Using LLMs. *Annual Meeting of the Association for Computational Linguistics*. <http://arxiv.org/pdf/2406.04482>

Karbasi, A., Montasser, O., Sous, J., & Velegkas, G. (2025). (Im)possibility of Automated Hallucination Detection in Large Language Models. *arXiv.org*. Unknown URL.

Kazlaris, I., Antoniou, E., Diamantaras, K., & Bratsas, C. (2025). From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs. *Applied Informatics*. Unknown URL.

Lee, M. (2023). A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics*. <https://www.mdpi.com/2227-7390/11/10/2320/pdf?version=1684231674>

Lee, D., & Yu, H. (2025). REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models. Unknown Venue. Unknown URL.

Li, H., Appleby, G., Alperin, K., Gomez, S., & Suh, A. (2025). Mitigating LLM Hallucinations with Knowledge Graphs: A Case Study. *arXiv.org*. Unknown URL.

Li, N., Li, Y., Liu, Y., Shi, L., Wang, K., & Wang, H. (2024). Drowzee: Metamorphic Testing for Fact-Conflicting Hallucination Detection in Large Language Models. Unknown Venue. <https://doi.org/10.1145/3689776>

Liang, X., Song, S., Niu, S., Li, Z., Xiong, F., Tang, B., Wy, Z., He, D., Cheng, P., Wang, Z., & Deng, H. (2023). UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation. *Annual Meeting of the Association for Computational Linguistics*. <https://arxiv.org/pdf/2311.15296>

Liu, Q., Chen, X., Ding, Y., Xu, S., Wu, S., & Wang, L. (2025). Attention-guided Self-reflection for Zero-shot Hallucination Detection in Large Language Models. *arXiv.org*. Unknown URL.

Munakata, S., Fukui, T., & Mohri, T. (2024). A Multiple-Fill-in-the-Blank Exam Approach for Enhancing Zero-Resource Hallucination Detection in Large Language Models. *arXiv.org*. Unknown URL.

Nishat, N., Coletta, A., Bellomarini, L., Amouzouvi, K., Lehmann, J., & Vahdati, S. (2025). Aligning Knowledge Graphs and Language Models for Factual Accuracy. *arXiv.org*. Unknown URL.

Qi, S., Cao, R., He, Y., & Yuan, Z. (2025). Evaluating LLMs’ Assessment of Mixed-Context Hallucination Through the Lens of Summarization. *Annual Meeting of the Association for Computational Linguistics*. Unknown URL.

Rabinovich, E., Ackerman, S., Raz, O., Farchi, E., & Anaby-Tavor, A. (2023). Predicting Question-Answering Performance of Large Language Models through Semantic Consistency. *IEEE Games Entertainment Media Conference*. Unknown URL.

Sansford, H., Richardson, N., Maretic, H., & Saada, J. (2024). GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework. Unknown Venue. Unknown URL.

Schmitt, R., Buelau, K., Martin, L., Huettl, C., Schirner, M., Stefanovski, L., & Ritter, P. (2025). Biological Database Mining for LLM-Driven Alzheimer’s Disease Drug Repurposing. *bioRxiv*. <https://www.biorxiv.org/content/biorxiv/early/2024/12/08/2024.12.04.626255>

Sharkey, E., & Treleaven, P. (2024). BERT vs GPT for financial engineering. *arXiv.org*. Unknown URL.

Silvestri, C., Roshal, J., Shah, M., Widmann, W., Townsend, C., Brian, R., LHuillier, J., Navarro, S., Lund, S., & Sathe, T. (2024). Evaluation of a Novel Large Language Model (LLM) Powered Chatbot for Oral-Boards Scenarios. *medRxiv*. <https://www.medrxiv.org/content/medrxiv/early/2024/06/01/2024.05.31.24308044.full.pdf>

Sun, Y., Yin, Z., Guo, Q., Wu, J., Qiu, X., & Zhao, H. (2024). Benchmarking Hallucination in Large Language Models Based on Unanswerable Math Word

Problem. *International Conference on Language Resources and Evaluation*. Unknown URL.

Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv.org*. Unknown URL.

Vladika, J., Soydemir, I., & Matthes, F. (2025). Correcting Hallucinations in News Summaries: Exploration of Self-Correcting LLM Methods with External Knowledge. Unknown Venue. Unknown URL.

Xie, Z. (2024). Order Matters in Hallucination: Reasoning Order as Benchmark and Reflexive Prompting for Large-Language-Models. *arXiv.org*. Unknown URL.

Yang, B., Mamun, M., Zhang, J., & Uddin, G. (2025). Hallucination Detection in Large Language Models with Metamorphic Relations. Unknown Venue. Unknown URL.

Zhang, H., Anjum, S., Fan, H., Zheng, W., Huang, Y., & Feng, Y. (2025). Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in Large Language Models. *arXiv.org*. Unknown URL.

Zhang, H., Zhang, J., & Wan, X. (2024). Quantity Matters: Towards Assessing and Mitigating Number Hallucination in Large Vision-Language Models. Unknown Venue. Unknown URL.

Zhang, C., Feng, Z., Zhang, Z., Qiang, J., Xu, G., & Li, Y. (2025). Is LLMs Hallucination Usable? LLM-based Negative Reasoning for Fake News Detection. *AAAI Conference on Artificial Intelligence*. Unknown URL.

Zhou, X., He, J., Ke, Y., Zhu, G., Guti'erez-Basulto, V., & Pan, J. (2024). An Empirical Study on Parameter-Efficient Fine-Tuning for MultiModal Large Language Models. *Annual Meeting of the Association for Computational Linguistics*. Unknown URL.