International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Hallucination and Bias Mitigation in Large Language Models: A Systematic Review of Self-Reflective and Drift-Aware Approaches

Deepak Kumar[1*(0009-0008-4956-142X)], Shilpa Suhag[2(0009-0002-9766-6345)]
[1]Research scholar, Department of Artificial Intelligence and Data Science (AIDS), Maharshi Dayanand University (MDU)
[2]Assistant Professor, Department of Computer Science (CSE), Maharshi Dayanand University (MDU),

Abstract - Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding, reasoning, and multimodal generation; however, hallucination and bias remain major challenges affecting their reliability and safe deployment. Based on the PRISMA guidelines, this review paper provides a systematic review of recent studies on hallucination detection, bias minimization, introspection, uncertainty quantification, and drift-aware modeling in LLMs and multimodal LLMs. Twenty-five studies were selected and analyzed from the major scientific literature databases. The review revealed common sources of hallucinations in multi-step reasoning: weak reasoning supervision, spurious correlation, semantic drift, lack of grounding, and error propagation. The results also demonstrated that self-reflective processes like self-consistency evaluation, reflective instruction tuning, self-verification, reinforcement learning from reflective feedback and uncertainty-aware introspection can substantially boost factual consistency and reasoning reliability. Furthermore, two frameworks that use drift awareness and topology-based reasoning showed great promise in detecting progressive reasoning deviations and epistemic gaps in the generation process. The reviewed works also showed that hallucination and bias matter more in domain specific use-cases, as well as safety-critical applications, such as healthcare, multimodal reasoning, recommender systems and decision-support applications. The findings of this study highlight the need for self-reflective reasoning, uncertainty awareness, and drift-monitoring mechanisms to ensure the robustness and trustworthiness of future systems based on LLM. The review also highlights areas where research has failed to meet the needs and outlines future research directions in the areas of adaptive, reliable and self-corrective language model architectures.

Keywords: Large Language Models, Hallucination Detection, Bias Mitigation, Self-Reflection, Semantic Drift, Drift-Aware Framework, AI Safety, Natural Language Processing

1. INTRODUCTION

Large Language Models (LLMs) are rapidly evolving as one of the most promising forms of AI that can handle a variety of NLP tasks such as question answering, summarization, dialog generation, reasoning, recommendation and decision support [1]. They have proven to be very versatile and are already in use in various fields including intelligent assistants, recommender systems, multimodal interactions, education, finance, and healthcare [2]. In recent years, the capabilities of multimodal large language models (LLM) have also been extended to include textual and visual reasoning in single generative models [3]. Even with these great strides, issues of reliability and trustworthiness of LLM outputs persist and the prevalence of hallucination and bias continue to be major concerns [4].

Hallucination is defined as producing outputs that are factually incorrect, misleading, ungrounded, or even fabricated, but seem to make sense in a linguistic sense [5]. Bias, however, is the reproduction or reinforcement of social, unfair, or statistically biased patterns present in training data, or as a result of alignment processes [6]. Previous research has demonstrated that errors, such as hallucination, can arise during multi-step reasoning from lack of reasoning, spurious correlations, lack of foundations, semantic drift, and progressive error propagation [7].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

These problems are especially significant in safety-critical applications like healthcare, mental health management systems, legal applications, recommendation systems, and autonomous decision-making applications, in which erroneous or biased results can cause serious problems [8].

In response to these challenges, various approaches that can help mitigate hallucinations and improve reliability have been investigated such as retrieval-augmented generation (RAG) [21] or prompt engineering [29],[30], uncertainty estimation [22] and reinforcement learning from human feedback (RLHF) [27] and reflective instruction tuning [18] and self-verification mechanisms [13] and post-hoc reasoning correction methods [9]. In particular, self-reflective reasoning is a recent promising direction to enhance factual consistency and reasoning reliability in LLMs. Self-reflective frameworks allow models to assess intermediate reasoning steps, detect inconsistencies, to measure uncertainty and modify output through internal feedback mechanisms without completely depending on external supervision [10].

Meanwhile, the importance of semantic and reasoning drift in hallucination propagation has been increasingly highlighted in recent research. In long iterations of model output, the deviation from the factual basis, the given instructions, or the logical reasoning, is called drift [11]. A number of studies indicated that hallucinations are not just a single generation error, but instead are progressive in nature, with errors in early reasoning stages cascading through subsequent ones, resulting in reasonable but incorrect conclusions [12]. Hence, drift-aware monitoring mechanisms and topology-based reasoning analysis have become a point of interest as a strategy to enhance hallucination detection and reasoning stability in large language models [13].

Besides, uncertainty-aware and self-corrective approaches have shown promising results in enhancing the transparency and reliability of LLM systems. Recent literature on reflective learning [15], entropy-aware introspection [16], confidence calibration [17] and reinforcement learning from reflective feedback [18] suggest that reasoning enhancement in conjunction with self-evaluation mechanisms can lead to a substantial reduction in the frequency of hallucinations and the grounding in factual knowledge. But, despite all these developments, many current methods focus on the problem of hallucination, bias, uncertainty estimation and reasoning drift as independent tasks instead of as closely interconnected components of the model reliability [15].

Therefore, this review paper presents a PRISMA-guided systematic review of recent studies related to hallucination detection, bias mitigation, self-reflective reasoning, uncertainty estimation, and drift-aware modeling in large language models and multimodal large language models. Twenty-five recent studies were systematically analysed to explore the main causes of hallucination, the efficacy of existing mitigation measures and the benefits of reflective reasoning for enhancing the faithfulness and reliability of factual consistency. Through literature review this study documents and outlines current approaches, identifies gaps in the research and identifies the significance of incorporating SRL, DA awareness and DA monitoring in future LLM architectures. The results will be used to develop more trustworthy, reliable, interpretable, and adaptive large language model systems for real-world applications.

2. REVIEW METHODOLOGY

2.1 Literature Search Strategy

The method of this review paper was systematic literature review, which was designed to find, assess and combine the latest research on hallucination detection, bias mitigation, self-reflective reasoning and drift-aware mechanisms in large language models (LLMs). To comprehensively cover the research on artificial intelligence, natural language processing, and multimodal learning, the literature search was carried out on major scientific databases, such as Scopus, Web of Science, ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar.

Targeted keywords and search strategies were used to retrieve relevant studies, including “LLM hallucination detection,” “bias mitigation in large language models,” “self-reflection in LLMs,” “semantic drift in language models,” “reflective instruction tuning,” “retrieval-augmented generation,” “reasoning hallucination,” “uncertainty estimation in LLMs,” “multimodal hallucination mitigation,” and “AI alignment and self-verification.” Boolean operators (AND, OR, NOT) were used to narrow the search and enhance the results. Both

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

forward and backward citation tracking were also conducted to find influential studies and highly cited studies that are related to hallucination, bias, and reflective reasoning mechanisms.

The preliminary search process around 420 papers published from 2023 to 2025, highlighting the fast-paced development of LLM research.

2.2 Inclusion and Exclusion Criteria

Specific inclusion and exclusion criteria were set to ensure the relevance, quality and technical soundness of the reviewed literature. The inclusion criteria included peer-reviewed journal articles, conference papers, workshop papers and quality high-quality preprints, where the focus was on detecting, mitigating, reducing bias, incorporating self-reflection components into learning, improving reasoning, estimating uncertainty and incorporating drift components into modelling in LLM and multimodal LLM. Experimental frameworks, reflective training strategies, retrieval enhanced systems, reinforcement learning–based alignment, and reasoning verification mechanisms were highlighted in the studies. The recent publications are highlighted and span from 2023 to 2025, to reflect the latest advances in LLM reliability and safety research.

Studies were removed if they were not about LLMs, the methodology was not clearly explained, the study didn't have enough experimental results, or it was only concerned with unrelated machine learning areas, without having to do with hallucination, bias, reasoning or reflective mechanisms. No papers published in languages other than English, no duplicate papers, no opinion papers, and no papers with unclear evaluation procedures, were included. Also, papers not directly related to general AI ethics were excluded where appropriate to maintain the consistency of the review.

2.3 Study Selection Process

The procedure for selecting studies was conducted following the PRISMA 2020 framework to ensure transparency, reproducibility and systematic search of the literature. In the identification phase, about 420 records were retrieved from the various academic databases. Following the removal of 95 duplicate records, there were 325 studies for title and abstract screening.

Studies not related to the halluciantion detection, bias mitigation, reflective reasoning, or LLM drift-aware modeling were not included during screening. This resulted in the removal of 250 records because they were not relevant to the aims of this review. The other 75 articles were then fully evaluated for eligibility for inclusion in full text.

Each article was evaluated at the eligibility stage based on methodological clarity, experimental validation, relevance to self-reflective and drift-aware frameworks and contribution to hallucination or bias mitigation research. Studies with less transparency of evaluation metrics, less reproducibility of the experiments, or studies that do not directly relate to LLM reliability were excluded. After this evaluation, 25 studies were chosen for qualitative synthesis and review.

The studies selected were then subdivided into thematic categories such as self-reflection frameworks, multimodal hallucination mitigation, reasoning enhancement, uncertainty estimation, reinforcement learning–based alignment, semantic drift analysis, and bias-aware generation techniques. These studies laid the groundwork for identifying existing problems, trends, and research gaps, which, in turn, contributed to the creation of the proposed drift-aware self-reflective framework for enhancing the reliability and trustworthiness of LLMs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

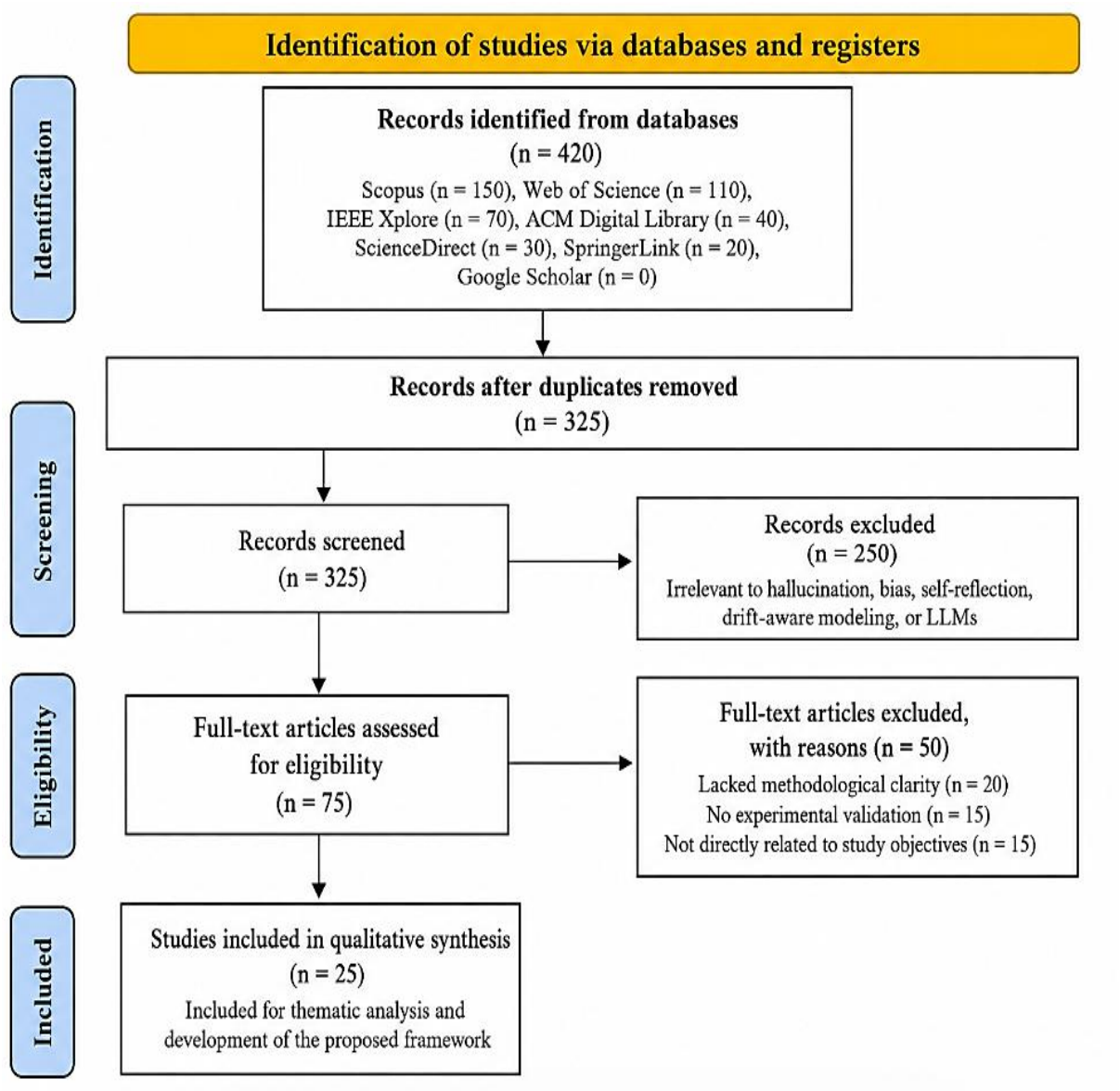

Figure 1. PRISMA Flow Diagram Illustrating the Study Selection Process for Systematic Review

3. FINDINGS AND DISCUSSION

3.1 Major Causes of Hallucination in Large Language Models

The reviewed studies all confirmed the existence of hallucination in large language models, revealing that it is a result of several interconnected factors, such as weak reasoning supervision, spurious correlations, insufficient grounding, and semantic drift during multi-step generation. They noted that due to lacking of intermediate reasoning supervision, large vision-language models often generate hallucinated results when they use simple correlations between instructions and visual inputs [12, 15]. Wang et al. [14] also noted text-visual bias and co-occurrence bias as key causes of hallucinations in multimodal systems, which rely heavily on text prompts or often on the same types of objects, but not necessarily on the visual evidence.

Wang et al. [16] also showed that the hallucination effect from false correlations in the training data is very confident and hard to detect through traditional methods of uncertainty estimation and/or confidence filtering. They found that even for advanced models, contamination of training data with statistical biases can negatively impact the reliability of hallucination detection. Furthermore, Zhang et al. [25] demonstrated that hallucinations from large reasoning models tend to arise from reasoning drift, where reasoning errors at the beginning of the chain of thought are amplified through a long reasoning process and integrated into coherent but wrong output.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

All of these results point to a non-accidental generation error as a structural problem linked to reasoning quality, data-bias and progressive semantic deviation, rather than a random generation error.

3.2 Role of Self-Reflection in Hallucination Mitigation

One of the common findings in the literature reviewed is the growing application of self-reflective reasoning mechanisms to enhance factual consistency and to minimize hallucinated output. Han et al. [11] introduced a self-consistency framework that matched long-form responses to short responses (binary) to detect inconsistencies and automatically generate preference pairs for training. Their results showed that internal consistency signals can effectively decrease hallucinations without affecting instruction-following performance.

In the same manner, Ji et al. [13] proposed an interactive self-reflection mechanism in a Q&A system that incorporated iterative feedback and knowledge acquisition by iteratively refining outputs with self-correction. They found that reflective evaluation helped to enhance consistency of facts and lessen hallucinated medical responses. Yan et al. [23] also added a self-verification mechanism to multimodal reasoning systems, allowing models to re-examine their prediction of an object and ground the visual prediction when making an inference.

Other studies highlighted the reflective learning in training. The reflective instruction tuning with rationale generation technique was found to substantially enhance reasoning capability and mitigate hallucinations in vision-language tasks by Zhang et al. [12] and Zhang et al. [15] respectively. Similarly, Lee et al. [34] suggested Reinforcement Learning from Reflective Feedback (RLRF) which models refined outputs with internal reflective evaluation prior to reinforcement learning optimization. Factual and downstream reasoning were enhanced in experimental groups.

3.3 Reasoning Enhancement and Uncertainty Awareness

This work emphasized the need for reasoning enhancement and uncertainty modelling to boost the reliability of LLM.Several studies pointed out the significance of reasoning enhancement and uncertainty modelling to enhance the reliability of LLM. Leite et al. [18] explored reasoning techniques like Chain Of Thought and Tree Of Thought prompting, demonstrating that deliberate reasoning techniques derived from cognitive science can enhance logical consistency and mitigate the generation of unreliable outputs. Li et al. [19] also pointed out that reasoning enhancement boosts the internal logical continuity, especially when paired with the retrieval-augmented generation systems.

Yan et al. [20] studied how self-reflection is limited in small models and introduced a Entropy-based Introspection framework, called Entrospect, which enhanced the reasoning quality and efficiency of small models. Xu et al. [28] also showed that self-reflective rationales and calibrated confidence estimation helps with uncertainty awareness and mitigates hallucinated responses with over-confidence.

Kirchhof et al. [30] and Kirchhof et al. [33] also investigated uncertainty representation in LLMs and concluded that current models are not very good at providing natural language explanations of internal uncertainty. Through their studies, they found that sampling-based reflective summarization methods more accurately generate the uncertainty representations than using only a simple confidence score and hedging phrase.

3.4 Drift-Aware and Knowledge-Gap-Oriented Frameworks

Semantic drift and epistemic gaps during generation was another key insight gained from the literature. The Empty Brain Hypothesis (EBH) of Thambi et al. [21] suggested that hallucination is the result of silent gaps in knowledge and semantic drift in model cognition. They presented several metrics that they believe indicate epistemic failures during generation, including perplexity spikes and Gap Score G(t), and they showed notable decreases in hallucination rates.

To further support the drift-aware analysis, Zhang et al. [25] modeled reasoning trajectories as topological graphs. They developed a graph-based approach for their framework, providing structural signatures connected to hallucinated reasoning paths, and demonstrating that reasoning drift can be captured by analyzing the graphs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

From these studies, it can be inferred that hallucination usually emerges over time in the process of reasoning, not as isolated output errors.

3.5 Domain-Specific Hallucination and Bias Challenges

The reviewed papers also emphasized the importance of hallucination and bias in domain-specific and safety-sensitive applications. In healthcare and mental health environments, where the consequences of hallucinations can be profound, Ji et al. [13] and Asha et al. [24] highlighted the dangers of false or misleading outputs affecting patient care, directing patients to the wrong treatment, or postponing professional help. They showed that they needed more robust factual knowledge, selective abstention models and domain-aware reasoning models.

In fact, it was Hallucination effects and biases have also been found to impact decision-support applications, such as those found in the design and planning process by Muhs et al. [26]. Despite strides in personalisation and semantic understanding, recommender systems driven by LLMs continue to suffer from problems such as hallucination, computational inefficiency, and biased recommendation patterns, as observed by Guan et al. [29].

The results of this study suggest that the development of hallucination mitigation strategies will need to take account of the application context, particularly in contexts where factual reliability, fairness and safety are important requirements.

3.6 Comparative Summary of Reviewed Studies

Table1. Comparative studies

Ref.	Main Focus	Proposed Approach	Key Contribution
[11]	Vision-language hallucination	Self-consistency self-reflection	Improved factual consistency without external supervision
[12], [15]	LVLM hallucination	Reflective instruction tuning	Enhanced reasoning supervision using rationales
[13]	Medical QA hallucination	Interactive self-reflection	Improved factual medical response generation
[14]	Multimodal hallucination	Gradient-based constrained decoding	Reduced visual hallucination and bias
[16]	Spurious correlation hallucination	Bias-oriented hallucination analysis	Showed limits of confidence-based detection
[17]	Hallucination taxonomy	Mitigation strategy classification	Structured overview of mitigation methods
[19]	RAG and reasoning	Integrated hallucination mitigation	Combined grounding and reasoning approaches
[20]	Small model reflection	Entropy-aware introspection	Improved reflection in resource-limited models
[21]	Semantic drift	Empty Brain Hypothesis	Drift-aware hallucination detection
[23]	Object hallucination	Self-verification framework	Improved multimodal factual grounding
[25]	Reasoning hallucination	Graph-based topology analysis	Detection of hallucinated reasoning paths

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[28]	Confidence estimation	Self-reflective rationales	Improved uncertainty calibration
[34]	RL-based reflection	Reflective feedback learning	Enhanced reasoning and factuality
[35]	Self-learning agents	Multi-level reflection synthesis	Improved reflective self-improvement

4. CONCLUSION

This review paper systematically analysed the latest research on hallucination detection, bias mitigation, self-reflective reasoning, uncertainty estimation and drift-aware mechanisms in LLM and MLLM. A systematic review of 25 recent studies were selected and synthesised through a PRISMA-guided systematic review methodology to investigate the key contributors to hallucination, and the effectiveness of current mitigation strategies. The literature reviewed showed that common causes of hallucination in LLMs are weak reasoning supervision, correlation issues in training data, lack of grounding, semantic drift and progressive error propagation in multi-step reasoning processes. Further, a number of studies demonstrated strong connection between hallucination and bias in domain-specific or safety-critical applications including healthcare, mental health systems, recommendation systems, and multimodal reasoning environments.

The study also revealed that self-reflective mechanisms are one of the most hopeful research avenues for enhancing factual consistency and reliability in reasoning from LLM systems. The methods that achieved the best results in reducing hallucinations and improving reasoning included self-consistency evaluation, reflective instruction tuning, self-verification, entropy-aware introspection, reinforcement learning with reflective feedback, and multi-level reflection synthesis. Likewise, the uncertainty-aware frameworks and the drift-aware reasoning analysis were demonstrated to be helpful for model calibration, identify epistemic gaps, and minimize the progressive reasoning deviations in the process of model generation.

Additionally, the reviewed studies identified several significant research challenges, such as the lack of reflective capabilities in smaller language models, challenges with computing uncertainty, the continued existence of spurious correlation–based hallucinations, and the need for standardized evaluation benchmarks for reflective reasoning systems. Moreover, current solutions tend to deal with individual mitigation methods and look for overall adaptive solutions that can continuously track reasoning consistency and modify the errors across the inferences.

REFERENCES

[1] X. Zou *et al.*, “Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models,” *ArXiv Prepr. ArXiv241003577*, 2024.
[2] E. M. Weisman, “Investigating Key Structures in Protective Scenes for LLMs,” 2025.
[3] C.-K. Chia, A. Rames, and A. Z. A. Razak, “AI Chatbots in Research: Yes or No? A Self-reflective Exploration.,” *Pertanika J. Sci. Technol.*, vol. 33, no. 1, 2025.
[4] X. Liu *et al.*, “Enhancing large language models with multimodality and knowledge graphs for hallucination-free open-set object recognition,” *Proc. VLDB Endow. ISSN*, vol. 2150, p. 8097, 2024.
[5] A. Diržytė, “Large language models and the enhancement of human cognition: Some theoretical insights,” *Filos. Sociol.*, vol. 36, no. 1, pp. 14–22, 2025.
[6] C. Zhang *et al.*, “Seeing is Believing? Mitigating OCR Hallucinations in Multimodal Large Language Models,” in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*,
[7] S. Suzuoki and K. Hatano, “Reducing hallucinations in large language models: A consensus voting approach using mixture of experts,” *Authorea Prepr.*, 2024.
[8] S. T. K. Tadala, “Structured reasoning with large language models,” *Preprints*, 2025.
[9] S. T. K. Tadala, “Structured reasoning with large language models,” *Preprints*, 2025.
[10] Z. He *et al.*, “Seeing is believing? mitigating ocr hallucinations in multimodal large language models,” *ArXiv Prepr. ArXiv250620168*, 2025.
[11] M. Han *et al.*, “Self-Consistency as a Free Lunch: Reducing Hallucinations in Vision-Language Models via Self-Reflection,” *ArXiv Prepr. ArXiv250923236*, 2025.
[12] J. Zhang, T. Wang, H. Zhang, P. Lu, and F. Zheng, “Reflective instruction tuning: Mitigating hallucinations in large vision-language models,” in *European Conference on Computer Vision*, Springer, 2024, pp. 196–213.
[13] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, “Towards mitigating LLM hallucination via self reflection,” in *Findings of the*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Association for Computational Linguistics: EMNLP 2023, 2023, pp. 1827–1843.

[14] S. Wang, M. Shen, N. Chang, C. Nguyen, H. Li, and J. M. Alvarez, "Mitigating Multimodal Hallucinations via Gradient-based Self-Reflection," *ArXiv Prepr. ArXiv250903113*, 2025.

[15] T. W. Jinrui Zhang, H. Zhang, P. Lu, and F. Zheng, "Reflective Instruction Tuning: Mitigating Hallucinations in Large Vision-Language Models".

[16] S. Wang *et al.*, "When Bias Pretends to Be Truth: How Spurious Correlations Undermine Hallucination Detection in LLMs," *ArXiv Prepr. ArXiv251107318*, 2025.

[17] I. Kazlaris, E. Antoniou, K. Diamantaras, and C. Bratsas, "From illusion to insight: A taxonomic survey of hallucination mitigation techniques in LLMs," *AI*, vol. 6, no. 10, p. 260, 2025.

[18] S. C. Bellini-Leite, "Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues," *Adapt. Behav.*, vol. 32, no. 4, pp. 329–343, 2024.

[19] Y. Li, X. Fu, G. Verma, P. Buitelaar, and M. Liu, "Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems," *ArXiv Prepr. ArXiv251024476*, 2025.

[20] T. Yan, Z. Lin, L. Zhang, Z. Sun, and Y. Gao, "Entrospect: Information-Theoretic Self-Reflection Elicits Better Response Refinement of Small Language Models," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 24563–24577.

[21] M. Thambi, "Silent Knowledge Gaps in Large Language Models and Pathways to Proactive Nourishment," 2025.

[22] R. Erbe, "Consciousness and AI: A Meta-Reflective Framework," *Available SSRN 5854902*, 2025.

[23] B. Yan and M. Fang, "ReSelfVerMM: mitigating hallucination in multimodal LLMs through dataset reconstruction and self-verification," in *Second International Conference on Image Processing and Artificial Intelligence (ICIPAI 2025)*, SPIE, 2025, pp. 307–312.

[24] N. E. J. Asha, "Mitigation of hallucination in response to mental health counseling by large language models," 2025.

[25] G. Zhang *et al.*, "Unraveling Hallucination in Large Reasoning Models: A Topological Perspective".

[26] N. Muhs and A. Stankowski, "Leveraging LLMs for Reflection : Approaches to Mitigate Assumptions within the Design Process," 2024.

[27] Z. Tao *et al.*, "A survey on self-evolution of large language models," *ArXiv Prepr. ArXiv240414387*, 2024.

[28] T. Xu *et al.*, "Sayself: Teaching llms to express confidence with self-reflective rationales," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 5985–5998.

[29] Z. Guan *et al.*, "Large Language Models for Recommender Systems: A Problem-Driven Survey," *Tsinghua Sci. Technol.*, 2025.

[30] M. Kirchhof, L. Füger, A. Golinski, E. G. Dhekane, A. Blaas, and S. Williamson, "Self-reflective Uncertainties: Do LLMs Know Their Internal Answer Distribution?," in *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

[31] A. Chegini *et al.*, "Reasoning's Razor: Reasoning Improves Accuracy but Can Hurt Recall at Critical Operating Points in Safety and Hallucination Detection," *ArXiv Prepr. ArXiv251021049*, 2025.

[32] Z. Ma, J. Liu, X. Luo, Z. Huang, Q. Zhu, and W. Che, "Advancing tool-augmented large language models via meta-verification and reflection learning," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 2078–2089.

[33] M. Kirchhof *et al.*, "SelfReflect: Can LLMs Communicate Their Internal Answer Distribution?," *ArXiv Prepr. ArXiv250520295*, 2025.

[34] K. Lee, D. Hwang, S. Park, Y. Jang, and M. Lee, "Reinforcement learning from reflective feedback (rlrf): Aligning and improving llms via fine-grained self-reflection," *ArXiv Prepr. ArXiv240314238*, 2024.

[35] Y. Ge, S. Romeo, J. Cai, M. Sunkara, and Y. Zhang, "Samule: Self-learning agents enhanced by multi-level reflection," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 16602–16621.