Available Online: 19 August 2026 DOI: 10.54254/2977-3903/2026.36207

Large Language Model self-reflection: advances, limitations and future directions

Zicheng Huang

Computer Science, New York University Shanghai Campus, Shanghai, China

zh3155@nyu.edu

Abstract. Large Language Model (LLM) has achieved great success in many fields over recent years, demonstrating a promising future. Nevertheless, hallucination remains a major impediment that limits the usability and reliability of LLM. In order to mitigate hallucination, researchers have proposed various approaches, among which self-reflection stands out. While the initial optimism about pure LLM self-reflection slowly fades away, its methodologies are becoming more complicated and interweaved with other approaches. Due to fact that the boundaries between self-reflection and many of its synonyms are ambiguous, and that researchers keep coining new concepts and terminologies that may overlap and intertwine intricately with each other, this paper aims to provide a clear definition of self-reflection and some frequently used terminologies. In addition, given that currently there is no systematic work sorting out papers in this field, this paper fills this gap by classifying different self-reflection methodologies and presenting paradigmatic researches in each category. By the end, this paper discusses the possible limitations and directions of future development in this field, predicting that self-reflection might evolve into a component in hybrid LLM training approaches that is critical to hallucination alleviation.

Keywords: Large Language Model, self-reflection, hallucination

1. Introduction

Recent years have witnessed the rapid development of Large Language Models (LLMs), which have achieved remarkable success in a wide range of fields and demonstrated significant proficiency when handling a great variety of tasks. Nevertheless, occasional but persistently existing hallucination, which is a phenomenon where LLMs generate unfaithful, irrelevant, or self-contradictory information, remains one of the most disastrous problems, posing a great threat to the credibility of LLMs and thus limiting their real-world applications [1]. For instance, in the field of education, students might be misled by LLM hallucination, given that they tend to be overly reliant on the LLMs and do not have the ability or the initiative to assess the information critically [2]. The issue can be more fatal in the medical domain, where LLMs can make erroneous diagnoses and recommend inappropriate treatments, leading to real-life risks [3]. In order to mitigate hallucination, researchers have not only utilized external resources, such as Retrieval-Augmented Generation (also known as RAG), but also proposed internal approaches aiming at leveraging LLMs' pre-trained knowledge and reasoning capabilities [4], among which self-reflection stands out as a new paradigm of

verbal reinforcement learning and a panacea that improves the performance of LLMs from every aspect [5]. Despite some researchers expressing skepticism in their work [6-8] about the prevalent optimism, which emphasizes that improvements achieved by self-reflection might be brittle and unrobust, LLM self-reflection remains a promising field awaiting further exploration.

Noticing a shortage of studies that systematically sort out state-of-the-art findings in this field, this paper aims to construct a comprehensive view of the current situation of LLM self-reflection. Specifically, by collecting, analyzing and evaluating high-quality papers and paradigmatic researches in the recent three years (2023-2026), this paper endeavors to discuss the definition and the boundaries of LLM self-reflection, present the current development in this field, and predict directions that may thrive in the future.

2. Definitions

While there is no universally acknowledged definition for self-reflection in the context of large language model, typically it refers to a metacognitive strategy and an introspective process where a large language model carefully examines its reasoning and output in order to identify and correct mistakes [9]. However, due to the lack of a clear definition, many researchers simply use self-reflection as another way of saying self-correction, self-verification, self-critique, self-evaluation, etc. For instance, when studying the limitations of LLMs' self-verification capabilities, Stechly et al. [8] distinguished the differences between self-verification and external verifications, but used the words "self-reflection", "self-verification" and "self-critique" interchangeably. To make the matter worse, researchers in this field keep coining new synonyms of self-reflection with ambiguous definitions or concepts that intertwine with self-reflection, including reflexion [5], self-refine [10], self-contrast [11], and ReVISE (Refine via Intrinsic Self-Verification) [12], to name but a few. In order to avoid confusion, it is high time we establish a more systematic and comprehensive view of these similar yet unidentical concepts.

While it is difficult to construct a precise literal definition for each of the terms, due to the fact that these concepts typically overlap and interweave in reality, it is much more intuitive and concise to classify them according to the functions they perform. In this way, we ingeniously circumvent the play on words and instead focus on the essence of the terminologies. Table 1 provides a concise and comprehensive comparison of some frequently used terms germane to self-reflection.

Table 1. A naïve understanding of similar concepts and their primary functions [9, 10]

Concepts	Naïve understanding (What it is)	Primary function (What it does)
Self-reflection	A process of LLM's introspection	Carefully examine the model's own reasoning and output in order to correct potential errors
Self-verification	A process of LLM verifying the correctness of its own output	Determine whether the output is correct on the model's own
Self-critique/ self-evaluation	A process of LLM evaluating its own reasoning and output, which is more likely to focus on the negative part	Identify the errors and weaknesses in the model's reasoning and output
Self-correction	A process of LLM correcting its own output	Revise the output to generate a better answer
Self-refine	An iterative process allowing LLM to improve the quality of output on its own	Improve the model's initial output through iterative feedback and refinement

Table 1 concludes that self-reflection is an all-encompassing term that can refer to all processes that involve a large language model introspecting its reasoning and output to improve its performance. In other words, the concept of self-reflection generally includes self-correction, self-verification and self-evaluation, while self-refinement can be seen as an iterative process comprised of multiple rounds of self-reflection, according to the definition proposed by Madaan et al. [10]. In this context, we can thus use "self-reflection" as an umbrella term, while viewing "self-correction, self-verification" and "self-critique" as functions or steps included in the process of self-reflection. It is also noteworthy that all of these processes do not necessarily emerge in a specific LLM self-reflection research, so it is appropriate to perceive any introspective process, regardless of what components it actually contains, as self-reflection, but not self-correction, self-verification or self-critique, if not contained.

3. Taxonomy

Once the meaning of self-reflection is determined, a systematic taxonomy to categorize and analyze relevant work can be established. As the methodologies of self-reflection continuously evolve, currently we can divide them into three categories according to the need for extra resources: naïve prompt-based self-reflection, fine-tuned self-reflective models, and scaffolding reflection.

3.1. Naïve prompt-based self-reflection

This is the primitive and primary stage of LLM self-reflection. In this stage, the methodology of self-reflection is rather naïve and simple: using prompts written in advance to instruct LLMs to reflect on their own chain of thought. This approach is based merely on preset prompts and involves neither extra model training nor external resources. One typical example of study using this approach is conducted by Renze and Guven [9], in which the researchers test the effectiveness of LLM self-reflection when providing LLMs with prompts containing hints to different extents. The paper claims that the more information the preset prompt includes, the better LLMs perform, and that even merely giving LLMs a second chance to solve the problems they previously made mistakes with can significantly improve LLMs' problem-solving skills [13]. Weng et al. [13], on the other hand, proposed a more innovative approach, which utilizes backward self-verification. Instead of directly evaluating the logic chain or verifying the answer, they generate multiple candidate answers, presume that one of these answers is correct each time, and infer whether this answer is consistent with the original conditions.

Despite the differences between forward reasoning and backward verification, both methods are merely prompt-based, thus confronted with severe instability. Weng et al. [13] admitted that their results rely heavily on the ability of LLM and the quality of manually constructed prompts, while Renze and Guven [9] noted that the difficulty of problems or tasks chosen also makes great impact, implying that the effect of merely prompt-based self-reflection might fluctuate significantly depending on various factors. Li et al. [6] even points out that while self-reflective prompting enhances LLMs' performance in TruthfulQA, it adversely affects results in HotpotQA, indicating extreme paucity of robustness in this approach.

3.2. Fine-tuned self-reflective models

In contrast with naïve prompt-based self-reflection, which is train-free, in this stage, reflection methodologies involve model fine-tuning. A paradigm of this approach is critique-in-the-loop self-improvement [14], which adds a fine-tuned critique model to supervise the reasoning model in the conventional self-improvement loop. Their model achieves consistent improvements without encountering a bottleneck like conventional self-

reflection models, and can enhance model's overall performance across all difficulty levels, showcasing incredible robustness. Another impressive work is PANEL [15], which stands for stepwise natural language self-critique, utilizing a fine-tuned model to generate human-like natural language critique to help enhance its own reasoning. Their model not only demonstrates a steady and significant improvement on computational accuracy compared to baseline, but also possesses greater scalability, due to the fact that only one model is being trained in their approach.

Nevertheless, this approach is also confronted with various challenges. Compared to train-free methods, fine-tuning self-reflection models typically consumes more computational resources and costs significantly more time, raising doubts about the feasibility of this approach when tackling more complex problems on a larger scale. In addition, the scarcity of deep theoretical understanding regarding to self-critique mechanisms prevents researchers from further optimizing the performance of these models [15] and alleviating hallucination that still occasionally appears in long thinking chains [14].

3.3. Reflection with external resources

While the previous two categories emphasize introspection, which forces LLMs to detect errors and refine outputs completely on their own, this type of approach takes advantage of external resources, which endow LLM with abundant reliable information and reduce the computational cost. SELF-RAG [16] exemplifies this hybrid method. Traditional Retrieval-Augmented Generation (RAG) methods often ameliorate hallucination at the expense of creativity and flexibility, while the improvements brought by pure introspection are often unsteady and uncontrollable, still submitting to factual errors or contextual inconsistencies from time to time. By ingeniously combining both methods, SELF-RAG trains LLMs to critically adduce evidence from external resources and evaluate the quality of the output comparing to the version without retrieval, absorbing the merits of both approaches while avoiding the flaws. Note that this work is different from Auto-RAG [17], which is also an autonomous retrieval-augmented model but does not involve any self-reflection processes.

Despite the incredible improvements in terms of performance, factuality and citation accuracy achieved by SELF-RAG, the researchers admit that outputs not fully supported by the citations still occasionally occur [16], suggesting that a reliable verifier might still be a necessity.

4. Future directions

As the methodologies of LLM self-reflection continue to evolve, the limitations of this approach also begin to emerge, albeit far from hitting a bottleneck. Currently, the following three directions show great promise and are in urgent need of further development.

4.1. Reflection with external resources

Recently, many researchers have thrown doubts on the prospect of pure introspective approaches [8, 18, 19], pointing out that factual errors may accumulate in the process of self-reflection if relevant knowledge is not included in the model's pre-training, and that the balance between computational complexity and robustness still remains tenuous. Therefore, combining self-reflection processes with reliable external resources might be hitherto the most promising approach. In addition to SELF-RAG [16], approaches such as self-reflective models equipped with a reliable external verifier [8] or combined with reinforcement learning [20] also achieve remarkable progress.

4.2. Prospective reflection

When handling an extremely complex conundrum, even a single minor mistake hidden in the reasoning can lead to a calamity, and it would be inefficient and ineffective to examine the whole reasoning chain all over again. In addition, sometimes the ramifications brought by the traditional self-reflective workflow, which is basically composed of trial-and-error and reflect, are unpredictable and irreversible. For example, too many failed attempts might contaminate the LLM context, hindering the quality of self-reflection. Hence, Hanyu Wang et al. [21] suggest that it is necessary for LLMs to think before act, proposing an innovative "PreFlect" strategy to let LLMs generate a plan and critique the plan before actual work. Their approach showcases significant improvements in tackling complex problems and remarkable transferability across different agent architectures, which is of great reference value and well worth further study.

4.3. Theoretical understanding of self-reflection mechanisms

Many researches, though equipped with compelling arguments and strong outcomes, cannot explain why each of the components in their approaches work in detail theoretically, revealing a great gap in theoretical study of advanced LLM self-reflection methods. Some researchers explicitly admit the flaws in their theoretical explanation [9] and call for more researches in this field [15], indicating that there is an urgent need for a deeper theoretical understanding of self-reflection mechanisms.

5. Conclusion

In summary, LLM self-reflection was once perceived as a panacea for hallucination, and is still a promising field contemporarily that is well worth further exploration, even though confronted with major challenges such as susceptibility to various factors and incapability of eliminating factual errors. Self-reflective approaches have demonstrated great adaptability and flexibility, benefiting remarkably from external resources and other LLM learning methods, such as deep learning and reinforcement learning. In this context, self-reflection might evolve into a generic module, which may play a crucial role in mitigating hallucination and avoiding meaningless trial-and-error.

Considering the rapid development of LLM and the tendency of self-reflection methodologies to grow increasingly interwoven with other approaches, it is admittedly of extreme difficulty to list all the researches and achievements in the field of LLM self-reflection. Therefore, instead of attempting to search for all relevant work, I carefully selected several paradigmatic researches and exciting findings in each section of this paper, which is bound to overlook researches that might be of equal importance. Nevertheless, I endeavor to establish a systematic and comprehensive view of LLM self-reflection and explain the advances, challenges and possible future development through analyzing and evaluating high-quality papers in this field. In the future, I will keep focusing on state-of-the-art findings in this field to ameliorate my review.

References

[1] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics, 51*(4), 1373–1418. https://doi.org/10.1162/coli_a_00516

[2] Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P. S., & Wen, Q. (2025). LLM agents for education: Advances and applications. In *Findings of the Association for Computational*

Linguistics: EMNLP 2025 (pp. 13782–13810). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.743

[3] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 314–334). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.21

[4] Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1827–1843). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.123

[5] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (Vol. 36)* (pp. 8634–8652). Curran Associates. https://doi.org/10.52202/075280-0377

[6] Li, Y., Yang, C., & Ettinger, A. (2024). When hindsight is not 20/20: Testing limits on reflective thinking in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3741–3753). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.237

[7] Hong, R., Zhang, H., Pang, X., Yu, D., & Zhang, C. (2024). A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 900–925). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.52

[8] Stechly, K., Valmeekam, K., & Kambhampati, S. (2025). On the self-verification limitations of large language models on reasoning and planning tasks. In *13th International Conference on Learning Representations, ICLR 2025* (pp. 8097–8150). International Conference on Learning Representations, ICLR.

[9] Renze, M., & Guven, E. (2024). *Self-reflection in LLM agents: Effects on problem-solving performance*. arXiv. https://doi.org/10.48550/arXiv.2405.06682

[10] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (Vol. 36)* (pp. 46534–46594). Curran Associates. https://doi.org/10.52202/075280-2019

[11] Zhang, W., Shen, Y., Wu, L., Peng, Q., Wang, J., Zhuang, Y., & Lu, W. (2024). Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3602–3622). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.197

[12] Lee, H., Oh, S., Kim, J., Shin, J., & Tack, J. (2025). *Revise: Learning to refine at test-time via intrinsic self-verification.* arXiv. https://doi.org/10.48550/arXiv.2502.14565

[13] Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., & Zhao, J. (2023). Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2550–2575). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.167

[14] Xi, Z., Yang, D., Huang, J., Tang, J., Li, G., Ding, Y., He, W., Hong, B., Do, S., Zhan, W., Wang, X., Zheng, R., Ji, T., Shi, X., Zhai, Y., Weng, R., Wang, J., Cai, X., Gui, T., ... Jiang, Y.-G. (2024). *Enhancing LLM reasoning via critique models with test-time and training-time supervision*. arXiv. https://doi.org/10.48550/arXiv.2411.16579

[15] Li, Y., Xu, J., Liang, T., Chen, X., He, Z., Liu, Q., Wang, R., Zhang, Z., Tu, Z., Mi, H., & Yu, D. (2025). *Dancing with critiques: Enhancing LLM reasoning with stepwise natural language self-critique*. arXiv. https://doi.org/10.48550/arXiv.2503.17363

[16] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *13th International Conference on Learning Representations, ICLR 2024* (pp. 9112–9141). International Conference on Learning Representations, ICLR.

[17] Yu, T., Zhang, S., & Feng, Y. (2024). *Auto-rag: Autonomous retrieval-augmented generation for large language models*. arXiv. https://doi.org/10.48550/arXiv.2411.19443

[18] Bilal, A., Mohsin, M. A., Umer, M., Bangash, M. A. K., & Jamshed, M. A. (2025). *Meta-thinking in LLMs via multi-agent reinforcement learning: A survey*. arXiv. https://doi.org/10.48550/arXiv.2504.14520

[19] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17889–17904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.992

[20] Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., Shen, H., Xiong, J., Xin, Y., Jiang, Y., Tao, C., He, Y., Zhang, M., & Yan, S. (2025). SRPO: Enhancing multimodal LLM reasoning via reflection-aware reinforcement learning. In *Advances in Neural Information Processing Systems (Vol. 38)*, pp. 153676–153713. Curran Associates. https://doi.org/10.52202/085713-5139

[21] Wang, H., Cao, Y., Lin, L., & Chen, J. (2026). *Preflect: From retrospective to prospective reflection in large language model agents*. arXiv. https://doi.org/10.48550/arXiv.2602.07187