Classification and Scenarios Analysis of Hallucination Suppression Methods Based on ReAct Framework of Reflection Mechanism

Ruiyi Wang

{wryxiariruqinqihua@gmail.com}

Hainan Bielefeld University of Applied Sciences, Danzhou, Hainan Province, 571700, China

Abstract. The illusion problem has become a major limitation that limits their reliability. The mechanism of reflection and ReAct framework give a valuable research path to enhance the reliability of model inference. The current study concentrates on illusion suppression in the system of ReAct in relation to the reflection mechanisms with the systematic review of the contemporary mainstream reflection techniques, alongside the comparative analysis of these techniques, in the light of illusion rate and accuracy. The findings demonstrate that the system of reflection could enhance the reliability of the inference and stability of the task execution in the ReAct system to a significant extent. Various reflection approaches have distinguished their merits in terms of performance enhancement and cost of computation where the hybrid reflection approach performs the most favorable in terms of combined performance when dealing with complex tasks.

Keywords: ReAct framework, reflection mechanism, hallucination suppression, method classification, adaptation scenario analysis

1 Introduction

The breakthrough development of Large Language Model (LLM), including chatgpt and deepseek, has seen development in natural language understanding, generation and reasoning tasks in recent years. The agent system recommends that the model has the capability of self-reasoning and dynamically means it to reason with the environment and to modify its behavior as it does its tasks unlike the traditional frameworks of a simple static text generation [1]. One of the commonest representatives of this concept is the ReAct framework [2]. This paradigm is a creative way of creating a micro-cycle of the three connections of thinking, action and observation. It involves the agent thinking before every action to make sense of where the current activity is, and where the next action is going, to seek feedback through interaction with the external environment or source of knowledge depending on action and process the feedback result through observation to inform the future reasoning, which thus increases the transparency, controllability and accuracy of decision making and capacity of LLM to accomplish complex logical tasks. It has set the paradigm of the agent infrastructure [2].

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365227

The ReAct framework is basically an enhancement of the model towards reasoning and coordination of actions during interactive activities. Nonetheless, it does not address the illusion issue on a fundamental level. The factual illusion issue of the concept of LLM is multiplied in the multi-step decision making method of the ReAct which results into the multiplicative error propagation, failure of carrying out the task and even crossing the security line, which practically limits the portability of the agents. The illusion issue in the ReAct framework should be chiefly manifested in several phases of the reasoning and interaction process, according to the current research. When the reasoning-execution loop structure is applied, errors in the model at one step would be directly carried over on the next step as inputs to the next decision, and errors would accumulate and multiply within the decision chain, greatly lowering the overall quality of the individual task. At the same time, in the process of interpreting the information of the environmental observation, the agent can develop inappropriate internal thoughts based on the semantic understanding bias or gaps in knowledge, thereby influencing future choices of actions. Moreover, the reasoning process in the model can be inconsistent in the multi-turn interaction or long-term task situations which can be in the form of logical inconsistencies, or discontinuities between cross-step reasoning. Such issues of illusions not only impair the performance of the task but can also introduce possible dangers of insecurity and of reliability in the implementation of practice.

In order to improve the process of self-reflection and correction of errors in the ReAct framework, scientists have suggested the introduction of a reflection mechanism into the decision-making and reasoning process. The generated output of the model can be checked and evaluated at certain nodes, and this reflection mechanism can help in quick detection of possible errors and initiating corrective or re-reasoning feedback mechanisms [3, 4, 5]. Particularly, the reasoning logic of the thinking phase, rationality of strategy employed in the action phase, and accuracy of information integration of the observation phase can be checked with help of the reflection mechanism.

In cases where the model is uncertain or has potential hallucination, reflection mechanism may assemble external resources with internal knowledge including invoking retrieval or triggering the use of a tool or a multi-path reasoning, which prevents the production and spread of hallucinations. Thus, reflection mechanism can be viewed as one of the technologies, which help to make the ReAct framework more reliable and secure. The objective of the given paper is to conduct a systematic study of the hallucination suppression technique in the framework of the reflection mechanism in the ReAct approach and evaluate the situations when they are applicable and the performance trade-offs. To begin with, the available reflection mechanisms are habitually categorized and examined. Second, a comparative analysis is done on the grounds of the hallucination rate and accuracy. Lastly, essential issues with complex task situations are outlined, and suggestions are made regarding the way forward in the future.

2 Method Classification

In this article, the researchers divide the methods of hallucination suppression into 2 dimensions, namely, the sources of information on which the reflection is based and the relative integration in the ReAct cycle. Finally, these are grouped into three types which include internal self-examination reflection, external auxiliary reflection and mixed reflection.

2.1 Internal self-examination and reflection methods

The essence of the internal self-checking and reflection process consists in entrenching the reflection steps in the thinkactionobservation cycle of ReAct and, thus, attaining the constant correction of the internal reasoning logic. Chain-of-Verification[3] is one of them, which follows the process of generating questions-answering-verifying-correcting to explicitly auto-check the model output and dynamically modify the reasoning process based on the outcome of the auto-checking process, and essentially address the illusion issue on the level of factuality and logic. Reflexion[4] presents the critique-correction process within the observation phase to make the model able to critique itself regarding the first answer it provides and detect the errors, ambiguities or uncertainty and produce an improved one in response, and repeat many times until the quality requirements are achieved. The rationality of the step is proved by Stepwise Reflection [5] that follows the Action in ReAct. When the check fails, it goes back to be a Thought stage so that it can make a decision once again. Self-Check[6] generates and compares findings of reasoning among various perspectives to identify possible internal contradictions in the reasoning process which suppresses the occurrence of illusions. In particular, this technique will produce at least two lines of reasoning based on different points of view on the same task in the thinking process and will test the consistency of facts and logic between them. As it is observed in Figure 1, the chain of reasoning is corrected on a continuous basis by conducting an internal reflection mechanism as the reasoning chain is constantly corrected by a cyclical output of verification and backtracking and thus enhances logical consistency and stability of the decisions made.

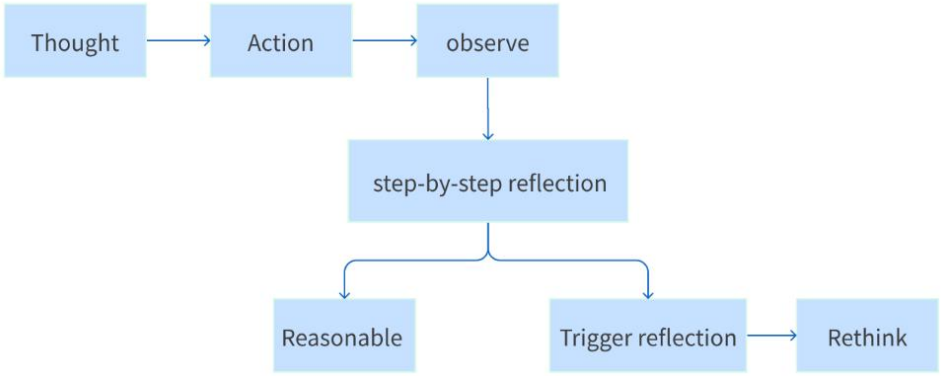

Figure 1. Flow chart of the entrapment process of the inner self-examination and reflection mechanism in the ReAct framework(Photo/Picture credit : Original).

The ReAct framework has the internal self-checking and reflection mechanism embedded in it, as shown in Figure 1. Having gone through the steps of Thought, Action and Observe, the model moves to progressive reflection module to measure the rationality of reasoning process at hand. In situations where the reasoning decision is found to be reasonable, the model proceeds with the execution of the next decision, when errors or uncertainties that might have arisen during the process are identified, the system initiates thinking and returns to the second stage of reasoning in order to repalan the decision course. The building presents the main concepts of internal consistency checking and cyclical correction.

2.2 External Auxiliary Reflection Methods

The effectiveness of ReAct thinking phase can be discussed through the introduction of external knowledge that prevents the cognitive flaws of the model. It operates on the idea of establishing a real-time knowledge retreiving and verifying and integrating closed loop in order to enhance the strength of the reasoning process [7,8]. Similar research shows that RAG [8] offers more factual endorsement of the model, which is to draw outside knowledge bases to the present task and the infusion of pertinent knowledge fragments into the process of thinking in reducing the factual illusion issue. Potential related review studies note that to a great extent, the factual reliability and reasonability of the model can be enhanced through the introduction of the retrieval models, tool calls and external memory modules. Meanwhile, there are works that get objective outputs calling external programs or interfaces to test the rationality of ReAct at the level of thoughts and actions in order to address the illusion of real-time due to the delay in information transmission or uncertainty [9]. Within this process such that the model determines that objective results are required to support the task, it will invoke the calls to tools and utilise the obtained results as a complement to the observation data, hence correcting the subsequent reasoning [10]. Moreover, the knowledge graph-enhanced reflective process cuts the reAct task into a path representation in an entity-relationship form and errors due to inappropriate association of information in the multi-step reasoning of the problem by the constraint of the reasoning direction with path constraints [11]. The proposed constraint mechanism is depicted in Figure 2 whereby the external knowledge query and logic verification modules are installed on the front end of the ReAct decision chain to minimise the risk of reasoning being off real-world knowledge.

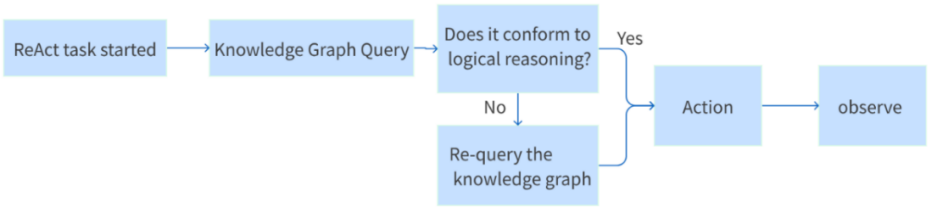

Figure 2. ReAct externally assisted reflection process schematic diagram was made using the external knowledge graphs(Photo/Picture credit : Original)..

Overall process of the process of the external auxiliary reflection in the framework of ReAct is depicted in Figure 2. The first stage of the knowledge graph is the task initiation, which initially queries the knowledge graph or external knowledge base and does a logical consistency check on the results of the search. Upon the conclusion of the logical reasoning constraints, the model passes to the Action and Observe phases, whereas in case of the uncertainty or contradicting results of the search, the system again asks the knowledge graph and recalculates the reasoning path. The reasoning effectively minimizes factual and relational errors on the reasoning in multi-step reasoning since the reasoning direction is limited by external knowledge.

2.3 Hybrid Reflection Method

Integrating the pros of self-checking and external auxiliary reflection, certain researches have been carried out to incorporate internal consistency checks and external verification systems into the ReAct framework to demonstrate synergistic inhibition of illusions of logic and illusions

of reality. On the more overarching process, the system initially depends on the internal reflection process to detect the possible error or uncertainties of the reasoning process and calls on external knowledge (to be recalled or environmental interaction) where required to do additional checking and correcting of the intermediate reasoning output, and thus, forms a closed-loop structure of identifying problems - verifying problems - correcting decisions. Regarding particular implementation, those methods generally introduce self-checking procedures like Reflexion or Stepwise Reflection into the original loop of ReAct to determine the rationality of the existing reasoning and actions; where the degree of uncertainty or illusion remains too high in the inner judgment of Reflexion, the system will also invoke the retrieval module or other tools to get objective information and feed it back in the next reasoning loop, thus enhancing the strength and effectiveness of the overall decision-making [12].

3 Application and Comparison of Evaluation Methods in Practice

Experimental Data set and validation objectives: This section will define the objectives of the experiment carried out to obtain good validation results, focused on the considerations made in this article.

3.1 Experimental Dataset and Validation Objectives

In this segment, the objectives of the experiment that will be performed to achieve good validation results will be established based on the considerations made in this article.
In order to assess the appropriateness of various reflection process to suppress hallucinations, Table 1 in this paper selected datasets representing different types of tasks to analyze them. The three methods (internal reflection, externally assisted reflection, and hybrid reflection) have the following data types, essentials, sample sizes, and validation goals, as illustrated in Table 1. As Table 1 reveals, internal reflection processes are those heavily dependent on the logical reasoning capacity of the model itself, external knowledge through real-time support is needed in external reflection processes, whereas the integration of the internal and the external reflection processes are needed in hybrid reflection processes.

Table 1. Dataset types and validation objectives adapted to different reflection mechanisms

Data types	Core features	Dataset	Sample size	Adaptation and verification target
internal	Relying solely on internal logical reasoning	TruthfulQA, HaluEval-Wild, StrategyQA	Approximately 38,600	Validating Illusion Repression through Internal Reflection Methods

external	It relies on real-time external knowledge and requires the use of external tools for support.	KILT-RAG, WebShop, WebArena-Lite	Approximately 445,087	Validating the suitability and effectiveness of external reflection methods
mix	It requires the collaboration of internal logic and external tools, integrating multimodal information such as text and environment.	ALFWorld, ScienceQA, WebArena, AgentBench （2024） MMMU, GAIA	Approximately 25,520	Verify the overall performance of the hybrid reflection method

Table 1 is a summary of different types of data sets and data validation goals of various reflection mechanisms. The internal datasets like TruthfulQA, HaluEval-Wild and StrategyQA are primarily considered to measure the capacity of the model to avoid illusion of pure logical reasoning, having a sample size of about 38, 600. Knowledge or tool interactions on the Web via real-time external knowledge are the basis of the External dataset like KILT-RAG, WebShop, and WebArena-Lite which use a sample size of about 445, 087 and are used to verify the flexibility and performance of the externally assisted reflection approach. Hybrid datasets like ALFWorld, ScienceQA, WebArena, AgentBench (2024), MMMU, and GAIA are a composite of text, environmental, and multi-modal data and are trained or tested to assess the overall performance of hybrid reflection processes in complex settings with a sample size of about 25, 520 and so on.

3.2 Evaluation Platform

vTo critically examine the suppression of hallucinations of varied reflection mechanisms using the ReAct framework, this paper may choose a few mainstream platforms to undergo experimental and analytical work. HELM offers a standardized large-model evaluation infrastructure, where repeatable and comparable evaluations of the performance of other models and methods under a common protocol and metric system can be performed. To quantitatively estimate the effect of reflection mechanisms on model performance, Hugging Face Evaluate aids in the determination of different metrics of automated evaluation, such as accuracy, robustness, and generation quality measurements. In evaluating the overall behaviour and generalisation capability of reflection mechanisms in realistic tasks of intelligent agents, AgentBench is concerned with complex task evaluation of intelligent agent systems, including multi-step inference, tool invocation and interaction with the environment, which are all necessary to assess the overall performance and generalisation capability of the mechanisms or methodology of reflection. Through the Synergism of these assessment media, the efficacy of the reflection mechanisms at three levels, namely standardized benchmark testing, an analysis of metric quantification, and verification of complex task situations can be thoroughly assessed in this paper.

3.3 Comparative Analysis Table of Different Reflection Methods

Three dimensions that include the rate at which the mechanisms of reflection suppress hallucination, and the level of accuracy, and when they can be applied are used to compare the differences in the performance of the mechanisms of reflections as discussed in this paper (Table 2). Table 2 displays methods of self-checking reflection, externally assisted reflection, and hybrid reflection in the works of representative methods. The hybrid reflection method has the best performance due to the table 2 results: best performance in the hallucination suppression and maximization of the accuracy.

Table 2. Comparison of the hallucination suppression effects and applicable scenarios of different reflection methods.

Method Category	Representative method	Hallucination rate suppression effect	Accuracy improvement (relative improvement)	Main applicable scenarios
Self-examination and reflection	Self-Check, Reflexion	middle	5%-15%	Logical reasoning, question and answer, multi-step reasoning
Self-examination and reflection	Gradual reflection	higher	10%-20%	Multi-step planning and interactive decision-making tasks
External Aid Reflection	RAG	high	15%-30%	Knowledge-intensive question answering and open-domain retrieval tasks
External Aid Reflection	Tool call verification	high	High (depends on the accuracy of the tool)	Search, computation, and environmental interaction tasks
Hybrid Reflection Approach	ReAct + Internal Reflection + External Tools Collaboration	Highest	More than 20%	Complex environments and multimodal tasks

Table 2 presents the comparisons of the various reflection techniques according to the illusion suppression effects as well as the effects of improving the accuracy. The findings indicate that

self-checking reflection mechanisms (Self-Check, Reflexion) moderate illusion suppression effect in logical reasoning tasks, and their improvement in accuracy is about 5 per cent -15 per cent; stepwise reflection mechanisms is better in multi-step planning and interactive decision-making tasks, whose improvement in accuracy is about 10 per cent -20 per cent. RAG is one of the highest illusion-suppression ability along with externally aided reflection, the accuracy scores in the method improve approximately 15-30 percent in the questions that are at knowledge level. The results of tool call verification is very high in terms of search, computation and environmental interaction test, however, it relies on the precision of external tools. By using an integration of not only the internal reflection method but also the external tool cooperation method, the hybrid reflection approaches have demonstrated the largest illusion suppression effect in more complex settings and in tasks with multimodal elements where the accuracy improves generally more than 20 percent, which is the current evolutionary trend in intelligent agent research. Moreover, knowledge editing techniques have been used to apply some modifications to the internal parameters of the model directly in order to rectify the bad factual knowledge as well as minimizing the likelihoods generated by the model parameter level level that would translate to some other potential underlying intervention path of the reflection mechanisms [13].

4 Future Outlook

The directions mentioned above are not free-standing but all, to a concerted effect, seek to ensure that we have a set of reflective agent systems that are efficient, regulable, and assessable. Despite the positive outcomes of the ReAct framework with reflection mechanisms, the eminent inapplicability of the tool in the stressful arsenal of a real-life setting remains an issue that needs more research and studying to be addressed. The results and discussion presented in this paper allow considering the following directions in the future of research concerning the ReAct framework using reflection mechanisms as its basis. First, the reflection triggering mechanism should change not only the representation of static strategies that depend on manually set rules, but also dynamic scheduling algorithms based on the estimation of uncertainties, error risks, or learning strategies, which should be more efficient and relevant. Second, the synergistic relationship between the depth of reflections, frequency of reflections and the overhead of inferences should be studied systematically to make sure that the hallucination suppression is effective as well as that the real-time needs of the real-life application are satisfied. Also, modeling of reflection mechanisms needs to be increased not only to pure-text-related tasks but also to multimodal-related tasks, dynamic interactions with the environment, and long-term planning of tasks to increase the generalization capacity and robustness of reflection mechanisms in real-world tasks. Lastly, a process-aware evaluation system needs to be built, to overcome the evaluation paradigm, which solely takes account of the correctness of the final outcome, and introduce process-level evaluation measures to the timing of reflection triggering, the efficacy of reflection and its cost, in order to gain a better picture on the actual worth of the reflection mechanism in the intelligent agent system. Generally, the future studies will advance reflection mechanism in an auxiliary error correction module to a main regulation and decision support element in the intelligent agent system, and the intelligent agent system would be developed to establish a highly reliable and secure general intelligent agent. Recent studies also present an intelligent agent architecture with evolvable memory system, which enhance the

ability of the intelligent agent on the task planning and decision making process in long term, and the evolution of long-term memory and experience gives a novel line of research in the co-evolution of the reflection mechanism and the memory [14].

5 Conclusion

The present paper is a systematic review and classification of existing mainstream approaches to suppressing illusion in the ReAct paradigm in terms of how it is reflected. Based on drawing comparisons of self-checking methods of reflection and the externally assisted reflection methods, the following conclusions can be made:
(1) Whereas reasoning-action loops are react-based, it is not enough to effectively suppress the illusion the reflection mechanisms are a powerful complement to enhance agent reliability.
(2) Self checking reflection methods are characterized by their high consistency in logic and coherence in their reasoning which are effective in eliminating the illusion of reasoning, they are weak in correcting errors of facts.
(3) Externally aided reflexive procedures wherein a retrieval or tool or structured knowledge is provided are very beneficial in the factual and real-time tasks but the complexity of its system and reliance on the resource is high.
(4) Hybrid reflection approaches combining internal and external verification performances in complex settings and multi-mod modal tasks are most effective when compared to other strategies, which corresponds to the present trend of developing agent systems.
Generally, various reflection mechanisms are bound to create tradeoffs among the effectiveness in suppressing illusions, improvement of accuracy, and system overhead; no universal solution can be said to be applicable to all cases.

References

[1] Schick, T., et al.: Toolformer: Language models can teach themselves to use tools. In: Advances in Neural Information Processing Systems (NeurIPS). 34 (2021)
[2] Yao, S., et al.: ReAct: Synergizing reasoning and acting in language models. In: Proceedings of the 39th International Conference on Machine Learning (ICML). pp. 22166–22188 (2022)
[3] Dhuliawala, S., et al.: Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv: 2309.11495 (2023)
[4] Shinn, N., et al.: Reflexion: Language agents with verbal reinforcement learning. arXiv preprint arXiv:2303.11366 (2023)
[5] Liu, X., et al.: Stepwise reflection for reliable agent decision-making. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1–15 (2024)
[6] Wang, Y., et al.: Multi-perspective reasoning for consistent answer generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). pp. 782–793 (2023)
[7] Nakano, R., et al.: WebGPT: Browser-assisted question-answering with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS). 34 (2021)
[8] Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9459–9474 (2020)

[9] Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (NeurIPS). 35 (2022)
[10] Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 11974–11984 (2022)
[11] Luo, L., et al.: Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. arXiv preprint arXiv:2410.13080 (2024)
[12] Meng, K., et al.: Locating and editing factual knowledge in GPT. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1739–1750 (2022)
[13] Mialon, G., et al.: Augmented language models: A survey. arXiv preprint arXiv:2302.07842 (2023)
[14] Zhang, Q., et al.: MemEvolve: Co-evolving memory systems for generalist agents. In: Proceedings of the 41st International Conference on Machine Learning (ICML). pp. 1–15 (2024)