Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28161

The Application of Reinforcement Learning in Reducing the Hallucination Problem of Large Language Models

Yuchen Liu

School of Information, Central University of Finance and Economics, Beijing, China
2024312298@email.cufe.edu.cn

Abstract. The paper is a review of the work regarding the use of reinforcement learning (RL) to solve issues of hallucinations in large language models (LLMs), providing a systematic review of mainstream methods, assessment standards, recent progress, and existing challenges. The proposed paper is devoted to such methodologies as prompt optimization, refinement of the reward mechanism, and enhancement of the quality of feedback, which all help improve the factual accuracy of model outputs by means of RL combined with human feedback and external validation. The evaluation criteria are divided into three dimensions: functional and factual consistency, alignment and reliability, specialized hallucination detection and assessment benchmark, and each of them is deeply discussed. Although optimized models have shown areas of notable improvement in particular tasks, serious difficulties have stayed, such as unequal assessment metrics, lack of benchmark coverage, incomplete correction of foundational model hallucinations, and a generalized applicability of optimization approaches. The focus of this review is to present insights that can be used in future studies in the field of RL-based methods to reduce the hallucinations in LLMs.

Keywords: Large Language Model, Reinforcement Learning, Hallucination Problem, Model Optimization, Evaluation Criterion

1. Introduction

In recent years, large language models such as Doubao and Deepseek have become increasingly popular and are being used more and more widely in daily life. They have brought a lot of convenience to people's study and life, and have also helped improve their work efficiency. However, large language models can also make mistakes in some cases, such as having hallucination problems, that is, the phenomenon of generating inaccurate, irrelevant or completely fabricated content. Moreover, large language models often have an extremely high level of confidence in such issues, which has also caused people's confusion.

To discuss these concerns, individuals at first categorized the hallucination troubles of large language models into different subcategories, including internal hallucinations where the produced content is in conflict with the given input context, and external hallucinations where the content produced can not be checked against the objective facts and real-world knowledge; or factual hallucinations where the content produced has errors in objective facts, data and events that are

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28161

inconsistent with the real world. They then delved into targeted solutions based on these classifications. A sequence of frameworks that can reduce the hallucination issues of large language models was built via machine learning or reinforcement learning, and more investigation was performed on these frameworks, including the objective of optimizing prompts within the framework to lead to the StablePrompt method [1].

This paper combs and analyzes the progress achieved in the field of training reinforcement learning to minimize hallucination issues in large language models to make a reference to future research. The review is an overview of certain mainstream strategies to minimize hallucination issues in large language models through reinforcement learning, the state of current evolution of the same and the criteria that are now as well as those that might be used to assess the problem of hallucination in large language models, the existing problems, and the potential solutions to them.

2. The application of reinforcement learning in reducing hallucinations in large language models

At present, there are various approaches to addressing the hallucination problem of large language models, mainly including prompt optimization, improvement of the reward mechanism, and enhancement of feedback quality.

In terms of prompt optimization, Minchan Kwon et al. focused on the optimization direction of prompt words in large language models and proposed the StablePrompt method based on reinforcement learning, as well as its extended version TTE-StablePrompt. The StablePrompt method defines prompt word optimization as an online on-policy RL problem. It achieves the optimization of prompt words through a closed-loop process: "agent LLM generates prompt words → target LLM outputs results → calculate rewards based on the results". To address the issue that a single prompt word is difficult to adapt to all input scenarios, an input-dependent prompt word generation method (TTE-StablePrompt) was extended. By incorporating the current input into the state and allowing the agent to generate exclusive prompt words for each input, the performance of complex tasks such as open-domain question answering and text summarization is significantly improved. Experimental results show that StablePrompt and TTE-StablePrompt both outperform existing methods (such as APE, ProTeGi, RLprompt), and are applicable to LLMs of different sizes (2B-13B) and types (Llama, Mistral, Gemma, etc.). They have for the first time achieved RL-based prompt word optimization for models larger than 7B [1].

With the issue of the reward mechanism improvement and re-training, Shen et al. established the contrastive rewards mechanism. This strategy will provide a better representation of the model by including penalty terms in the reward model to favour real information, making the model provide better output and minimizes the hallucination phenomenon, thus effectively reducing the hallucination phenomenon. Such reward optimization strategy has been utilized in the area of reinforcement learning from human feedback (RLHF) and to a great extent, the model has been made much more compliant with human preferences and indirectly, less likely to provide hallucinations [2]. Sun et al. suggested the Uncertainty Gradient Data Augmentation (UGDA) technique to resolve the one of the noise and uncertainty of feedback. This approach is effective by the fact that the policy re-trainings might be used to maintain the performance of the reward model in the distribution and also by enhancing the performance of the model to minimize hallucination [3].

As far as improving the quality of feedback is concerned, Wang et al. integrated the feedback of the verification test bench, and took the verification information as a key part of the feedback signal in reinforcement learning training for Verilog code generation in such a way that the generated

DOI: 10.54254/2755-2721/2025.LD28161

content was right, and any mistakes were avoided because of the hallucinations [4]. Besides, Gao et al. applied the Retrieval-Augmented Generation (RAG) method, combining the inherent knowledge of the large language model with external retrieved knowledge, contrasting and verifying the preliminary responses of large language models with external retrieved knowledge to ensure the accuracy of feedback content and finally ameliorates the hallucination issues of the large language model [5].

3. The mainstream evaluation criteria for the hallucination problem of large language models

There are various mainstream evaluation criteria for the hallucination problem of large language models. This article will provide a detailed overview of three aspects: the consistency between function and facts, alignment and reliability, as well as specific hallucination detection and evaluation benchmarks.

3.1. Dimension of consistency between function and fact

In the code generation scenario, with "functional correctness" as the core, the "pass@k" metric is used to measure the proportion of the generated code that passes the testbench verification in k attempts (k usually takes 1, 5, 10 in the field). The higher the proportion, the fewer hallucinations (functional errors) there are. Specifically, it is manifested in four core aspects: whether the code can pass compilation, whether it meets the requirements of test cases, whether the generated testbench matches the functionality of the code as expected (such as the consistency between the output and the actual output), and whether it complies with design specifications such as signal processing and temporal logic correctness. In the general scenario, the focus is on the matching degree of the generated content with the facts or the context. The core indicators include answer fidelity (whether the generated content is strictly based on the retrieved knowledge or the context, without fabricating without basis); answer relevance (whether it directly responds to the core of the question, avoiding irrelevant information); negative example rejection ability (whether it can reject the answer instead of fabricating when the retrieval results have no valid knowledge); and robustness (including noise robustness (accuracy when facing irrelevant documents) and counterfactual robustness (the ability to identify and ignore incorrect information)).

3.2. Alignment and reliability dimensions

In the alignment and reliability dimensions, they are further divided into three indicators: the scoring consistency of the reward model, the quality of the generated content, and the performance of adversarial tests. In the aspect of the scoring consistency of the reward model, it is measured by the variance of the scores given by the reward model to the generated content. High variance indicates an unreliable judgment of the content's authenticity, suggesting the possibility of hallucinations. In the aspect of generated content quality, it is evaluated through human assessment, GPT-4 judgment, and scores from third-party reward models (such as UltraRM, PairRM), to determine whether the generated content meets standards such as "helpful, harmless, and real". In the aspect of adversarial test performance, in the "escape" task (such as RED-EVAL), the probability of the model being induced to generate inappropriate content is measured by the attack success rate (ASR). The lower the ASR, the stronger the ability to resist hallucinatory output.

DOI: 10.54254/2755-2721/2025.LD28161

3.3. Specialized hallucination detection and assessment benchmark

In the specialized hallucination detection and evaluation benchmark summarized by Huang et al., it is divided into two aspects: hallucination detection benchmark and hallucination evaluation benchmark [6].

3.3.1. Hallucination detection benchmark

The hallucination detection benchmark is used to evaluate the performance of hallucination detection methods and verify their ability to identify false or inconsistent content. It mainly includes the following types:

Firstly, there are benchmarks in the biography and summary domains. For example, SelfCheckGPT-Wikibio [7], a synthetic Wikipedia biography generated based on GPT-3, with factual annotations, is used to detect hallucinations in biographical text; LSu [8] focuses on the summary task and evaluates the factual consistency between the summary content and the original text [8].

General and task-specific benchmarks: such as HaluEval [9], combining 5000 general user queries and 30000 task-specific samples, constructed through the "sampling - filtering" method, to evaluate the model's ability to identify hallucinations in different scenarios; BAMBOO [10] extends to long text detection, including long text hallucination samples related to scientific papers [10].

Secondly, there are cross-domain and multilingual benchmarks. Such as FELM [11], covering multiple domains such as world knowledge, science, and mathematics, with 3948 labeled samples, to evaluate the applicability of detection methods across domains; ChineseFactEval [12] is for Chinese LLMs, testing the detection ability of Chinese factual content [12].

Open-domain question-answering benchmarks: such as RealHall [13], SAC^3 [13], based on open-domain question-answering scenarios, to evaluate the detection effect of hallucinations in question-answer content; PHD [14] starts from Wikipedia entities and focuses on the detection of paragraph-level non-factual content [14].

3.3.2. Hallucination assessment criteria

The hallucination assessment benchmark aims to quantify the tendency of large language models (LLMs) to generate hallucinations, with a focus on factual accuracy and consistency with the context. It mainly includes the following types:

Long-tail factual knowledge benchmark: such as PopQA [15], Head-to-Tail [16], which is constructed based on the frequency of entities' appearance in the training data or domain-specificity, used to evaluate the model's mastery of low-frequency or specialized domain knowledge, often exposing the model's hallucinations in long-tail knowledge [15, 16].

First is the timeliness knowledge benchmark. Such as REALTIMEQA [17], FreshQA [18], which includes questions updated in real time (such as current events, dynamic knowledge), testing the model's ability to capture the latest information, revealing hallucinations caused by outdated knowledge [17, 18].

Second is the imitation false knowledge benchmark. Such as TruthfulQA [19], HalluQA [20], constructed through adversarial methods with questions that are prone to cause model errors (such as common misconceptions in the training data), evaluate whether the model will imitate false information to generate content [19, 20].

DOI: 10.54254/2755-2721/2025.LD28161

Third is the domain-specific benchmark. Such as Med-HALT [21], focusing on professional fields such as medicine, through domain-specific questions to evaluate the model's hallucinations in professional knowledge [21]; HaluEval 2.0 [22] covers multiple domains such as biomedicine, finance, and science, including 8,770 questions that are prone to cause hallucinations [22].

4. The limitations of current methods and future prospects

4.1. Current optimization effects of RL-based methods

In the VeriPrefer method, significant improvement in functional correctness was achieved through the automatic generation of test benches + DPO reinforcement learning: in the RTLLM v2 benchmark, the functional pass@1 reached 52.4%, pass@5 reached 66.4%, outperforming methods based on BLEU (48.0%) and AST (49.2%); in the VerilogEval v2 benchmark, the functional pass@1 reached 49.3%, pass@5 reached 64.7%, outperforming SFT (47.8%) and PPO (47.0%); After optimizing Qwen2.5-Coder-14B, the functional pass@1 in the VerilogEval-Human reached 61.9%, and cross-model application (such as RTLCoder) increased pass@5 by an average of 2.3%-2.7% [4].

In the comparison of reward methods, reducing hallucination through penalizing uncertainty was demonstrated: in human evaluation, the success rate in the harmless task of Anthropic/HH-RLHF was 63.7% (advantage difference 53.9), far exceeding SFT (9.8%) and PPO (14.7%); the success rate in the helpful task was 66.7% (advantage difference 45.1), with content more in line with human preferences;In GPT-4 evaluation, the same task success rate was 57.9% (advantage difference 50.1), reliability was improved; verification by third-party reward models (such as UltraRM) showed that the task success rate of Anthropic/HH-RLHF was 75.2%, and the task success rate of OpenAI/Summary was 97.5%; in adversarial testing (RED-EVAL), the average ASR of Mistral-7B-CR model was only 0.043, much lower than SFT (0.472) and DPO (0.161) [2].

In the RAG method, reducing hallucination through external knowledge retrieval, the fidelity of answers increased by 15%-30%, in benchmarks such as HotpotQA and NQ, models based on Wikipedia retrieval improved the fidelity in TriviaQA from 55% to 75%; noise robustness was enhanced, with an accuracy rate still above 70% when containing 30% irrelevant documents, while pure LLMs dropped below 40%; the negative example rejection rate reached 65%, much higher than that of pure LLMs (20% based on the RGB benchmark) [5].

4.2. Core existing problems

4.2.1. Assessment system problems

The first issue is the lack of uniformity in the detection and evaluation criteria for hallucinations. In the works of Wang et al. [4], Shen et al. [2], and Huang et al. [6], there are different criteria for evaluating the level of hallucinations, which may lead to the inability to compare the results of different standards and make it difficult to determine the actual ability of the model and framework to reduce hallucinations.

The second issue is the insufficient coverage of the detection and evaluation benchmarks. According to the specialized hallucination detection and evaluation benchmarks in Section 3.3, it can be observed that although the existing hallucination detection benchmarks cover biographies, summaries, common scenarios, etc., there are relatively few samples for emerging fields (such as real-time social content, cross-modal generation). The benchmarks for long-text hallucination detection (such as extremely long scientific papers, multi-document fusion generation) are still

DOI: 10.54254/2755-2721/2025.LD28161

limited (only a few benchmarks like BAMBOO involve them); the coverage depth of the evaluation benchmarks in terms of timely knowledge (such as the latest policies, sudden current events) and extreme long-tail knowledge (such as low-frequency knowledge in niche professional fields) is insufficient, making it difficult to fully expose the hallucinations caused by the model's knowledge lag or scarcity.

4.2.2. Foundational model problems

The third issue is that the hallucination problem of the basic model remains prominent. According to the article by Wang et al. [4], the answer fidelity of general LLMs in open-domain question answering is usually below 60% [4], and 30%-40% of the content in long text generation contains unverified errors [4]; for code-specific LLMs (such as CodeLlama 7B), the pass@1 of functional correctness is only 17.9% [4], even for the optimized model (such as Qwen2.5-Coder-14B after VeriPrefer optimization), the pass@1 is only 61.9%, and there is still a significant functional deviation [4].

4.2.3. Optimization method problems

The fourth point is that the limitations of the optimization methods are obvious. According to the article by Gao et al., although the VeriPrefer code optimization methods improve the functional correctness, the improvement in cross-model universality is limited (the average improvement in pass@5 is only 2.3%-2.7% for cross-model application such as RTLCoder) [4], and the hallucination reduction effect for complex design tasks (such as multi-module sequential logic) is still not satisfactory; The RAG method relies on retrieval quality. When the retrieval results contain noise (such as 30% irrelevant documents), although the model accuracy exceeds 70%, it still decreases significantly compared to the noise-free scenario, and it is still prone to fabricating content in the "knowledge blind areas" that are not covered by the retrieval [5]; Compared with the alignment methods such as contrastive rewards, although they reduced ASR to 0.043 in adversarial tests (such as RED-EVAL), they did not completely eliminate the possibility of being induced to generate hallucinations, and the adaptability in small language or cultural scenarios has not been verified [2]. Besides, although Sun et al. constructed a framework to reduce the hallucination problem of large language models through reinforcement learning, there are still problems in finding more precise rewards and tuning methods, as well as theoretical analysis of large language models [23]. The Retrieval-Augmented Generation technology, by combining reinforcement learning, helps to alleviate the hallucination phenomenon of the model, but still has certain difficulties when dealing with unfamiliar data [5].

4.2.4. Reward model problems

The fifth issue is that the reliability of the reward model and evaluation needs to be improved. According to the article by Shen et al., there is variance in the scoring of generated content by the reward model (high variance indicates unreliable judgment). The subjectivity of third-party evaluations (such as human evaluation, GPT-4 evaluation) may affect the consistency of the results, leading to deviations in the judgment of whether "hallucinations have decreased" [2]. In addition, although Shen et al. further improved reinforcement learning from human feedback (RLHF) and thereby reduced the hallucination problem, they still faced challenges such as reward model noise, sampling cost, and complex task adaptation [2].

DOI: 10.54254/2755-2721/2025.LD28161

4.2.5. Possible solutions to the problem

Regarding the lack of uniformity in the criteria for the hallucination problem of the large language model, it is necessary to identify the common points and different bias points among these criteria, combine them into a comprehensive framework for evaluating the hallucination problem of large language models, thereby establishing a unified evaluation standard. For other issues related to the reduction of hallucinations and the inadequacies of frameworks and models, further research can be conducted to expand the breadth and depth of the applicability of the frameworks and models, strengthen their integration with other related methods, and continuously optimize the models, thereby construct more effective frameworks and methods for reducing the hallucination problem of LLMs.

5. Conclusion

Reinforcement learning provides an effective approach to reducing the hallucination problems of large language models. Through methods such as prompt optimization, reward mechanism improvement, and feedback quality enhancement, it has achieved phased results in tasks like code generation, general question answering, and safety alignment. Among them, Prompt optimization (StablePrompt) is lightweight and efficient, suitable for specific tasks of small and medium-sized LLMs; reward mechanism improvement (contrastive rewards) can improve the alignment with human preferences, suitable for safety alignment tasks; feedback quality enhancement (RAG + test bench) can improve factual accuracy, suitable for code generation and open-domain question answering. The progress mentioned in the above article indicates that reinforcement learning can effectively alleviate hallucination problems by dynamically optimizing the model's behavior.

However, current research still faces significant challenges, such as inconsistent evaluation criteria. This article also sorts out and summarizes these issues. For future research, three aspects should be focused on: First, establish a unified hallucination evaluation standard and integrate multiple dimensions of indicators to form a comprehensive assessment framework; second, expand the coverage of detection benchmarks to include more domains and scenarios of samples; third, deepen the combination of reinforcement learning with external knowledge and multimodal feedback to enhance the model's anti-hallucination ability in complex tasks and unknown environments. Through continuous optimization of methods and benchmarks, reinforcement learning is expected to further drive the development of LLMs towards a more reliable, factual and trustworthy direction.

References

[1] Kwon, M., Kim, G., Kim, J., Lee, H., & Kim, J. (2024). StablePrompt: automatic prompt tuning using reinforcement learning for large language models. *arXiv preprint arXiv: 2410.07652*.

[2] Shen, W., Zhang, X., Yao, Y., Zheng, R., Guo, H., & Liu, Y. (2024). Improving reinforcement learning from human feedback using contrastive rewards. *arXiv preprint arXiv: 2403.07708*.

[3] Sun, Z., Guo, Y., Lin, Y., Chen, X., Qi, Q., Tang, X., & Wen, J. R. (2025). Uncertainty and influence aware reward model refinement for reinforcement learning from human feedback. In *The Thirteenth International Conference on Learning Representations*.

[4] Wang, N., Yao, B., Zhou, J., Hu, Y., Wang, X., Guan, N., & Jiang, Z. (2025). Insights from verification: Training a verilog generation LLM with reinforcement learning with testbench feedback. *arXiv preprint arXiv: 2504.15804*.

[5] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv: 2312.10997, 2*(1).

DOI: 10.54254/2755-2721/2025.LD28161

[6] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1-55.
[7] Miao, N., Teh, Y. W., & Rainforth, T. (2023). Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv: 2308.00436*.
[8] Feng, H., Fan, Y., Liu, X., Lin, T. E., Yao, Z., Wu, Y., ... & Ma, Q. (2023). Improving factual consistency of text summarization by adversarially decoupling comprehension and embellishment abilities of llms. *arXiv preprint arXiv: 2310.19347*.
[9] Li, J., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2023). Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv: 2305.11747*.
[10] Dong, Z., Tang, T., Li, J., Zhao, W. X., & Wen, J. R. (2023). Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv: 2309.13345*.
[11] Chen, S., Zhao, Y., Zhang, J., Chern, I.-C., Gao, S., Liu, P., & He, J. (2023). FELM: Benchmarking Factuality Evaluation of Large Language Models. ArXiv preprint abs/*2310.00741*. https: //arxiv.org/abs/2310.00741
[12] Wang, B., Chern, E., & Liu, P. (2023). Chinesefacteval: A factuality benchmark for chinese llms.
[13] Friel, R., & Sanyal, A. (2023). Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv: 2310.18344*.
[14] Zhang, J., Li, Z., Das, K., Malin, B. A., & Kumar, S. (2023). SAC3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *arXiv preprint arXiv: 2311.01740*.
[15] Yang, S., Sun, R., & Wan, X. (2023). A new benchmark and reverse validation method for passage-level hallucination detection. *arXiv preprint arXiv: 2310.06498*.
[16] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2022). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv: 2212.10511*.
[17] Sun, K., Xu, Y. E., Zha, H., Liu, Y., & Dong, X. L. (2023). Head-to-tail: How knowledgeable are large language models (LLMs)? AKA will LLMs replace knowledge graphs?. *arXiv preprint arXiv: 2308.10168*.
[18] Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., ... & Inui, K. (2023). Realtime qa: What's the answer right now?. *Advances in neural information processing systems, 36*, 49025-49043.
[19] Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., ... & Luong, T. (2023). Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv: 2310.03214*.
[20] Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv: 2109.07958*.
[21] Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., ... & Qiu, X. (2023). Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv: 2310.03368*.
[22] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv: 2307.15343*.
[23] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv: 2401.03205*.