Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33183

The Development and Application of Optimization Methods for Large Language Models Based on Reinforcement Learning

Putian Zhu

School of Software, Shandong University, Jinan, China
202400300323@mail.sdu.edu.cn

Abstract. With the wide application of large language models, their performance optimization has attracted more and more attention. In order to meet the challenges of the model in reasoning ability, alignment effect and generation quality, the large language model optimization method based on reinforcement learning has gradually become a research hot spot. This paper systematically analyzes the classification of reinforcement learning based large language model optimization methods in terms of reward mechanism and strategy optimization algorithm, as well as the introduction and analysis of various practical applications based on this, covering four areas: information retrieval and question answering, text semantic organization and generation, recommendation system and graph privacy protection. It is concluded that the current reinforcement learning-based large language model optimization methods have achieved significant results, but there are still some problems to be solved, such as data sets, black box effect and so on, which need further research and reflection in the future.

Keywords: large language model, reinforcement learning, reward mechanism, strategy optimization

1. Introduction

Nowadays, there are many large language models in the world, such as GPT-4, Llama, etc. They have achieved remarkable results in text generation, machine translation, question answering systems, and other tasks, which is due to their huge parameter scale and massive pre-training data. But they also have shortcomings, that is, when dealing with some problems with complex logical relations, they often give incorrect answers, and can not deal with the logical relations between them well, and often have problems such as logical chain fracture, factual errors, reasoning path confusion and so on. In order to deal with these shortcomings, scholars continue to research and explore, and find that the optimization method of large language model based on reinforcement learning is the most efficient and appropriate. Reinforcement learning is an excellent machine learning method that uses agents to obtain feedback and optimize behavior strategies by interacting with the environment. In recent years, reinforcement learning has been applied as the core idea in many research on the

© 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: 10.54254/2753-8818/2026.33183

optimization methods of reasoning ability of large language models, covering many levels, from reward mechanism to strategy update.

This paper first reviews the existing mainstream research methods based on reinforcement learning, and then based on the specific application of the current reinforcement learning based large language model optimization method in four areas, including information retrieval and question answering, text semantic organization and generation, recommendation system and graph privacy protection, and makes a summary and prospect reflection according to the progress achieved and the practical problems still faced.

2. Key optimization methods

Reinforcement learning is a machine learning method that relies on the interaction between agents and the environment and the reward mechanism to continuously learn. Its core lies in the environment and the reward mechanism. Therefore, different reward mechanisms and different strategic decisions will affect whether a large language model can adjust itself well in reinforcement learning to varying degrees.

2.1. Design of reward mechanism

At present, the mainstream of the design of reward mechanisms is divided into three types: rule-based reward, reward based on human feedback, and multi-dimensional composite reward. The rule-based reward mainly takes the accuracy of the results as the main reward mechanism for evaluation. Its characteristics are efficient calculation and strong explanation, but its accuracy also means that experts in this field need to give enough and accurate knowledge, and it is difficult to adapt to the complex and changeable application environment. Rewards based on human feedback are mainly constructed by collecting human scores or preferences for reasoning results. Typical methods are Reinforcement Learning from Human Feedback (RLHF) [1]. This method can better align human cognition, but it has problems such as high labor costs and difficult to ensure the consistency of feedback. The multi-dimensional composite reward needs to comprehensively consider the accuracy, logic, conciseness and other dimensions of the reasoning results, and then construct the composite reward function by weighting. This method can improve the comprehensive performance of model reasoning, but the determination of weights needs a lot of experimental tuning.

2.2. Improvement of the strategy optimization algorithm

Compared with the traditional reinforcement learning method based on the value function, the main advantage of the strategy optimization algorithm is that it limits the step size of the learning process, making the reinforcement learning process more coherent and stable, and it can naturally adapt to the continuous action space. Therefore, it is widely used in the fields of robot control and automatic driving. At present, the most popular strategy optimization algorithm is near Proximal Policy Optimization(PPO) [2]. With the advantages of good stability and high sample efficiency, it is frequently used in the process of reinforcement learning optimization. Such optimization methods as small batch updating and hybrid precision training can ensure the stability of updating while reducing the consumption of computing resources. In addition, there are hierarchical strategy optimization and offline reinforcement learning applications. The main idea is to divide the problem into many sub-problems and use the existing data for offline processing. They all have their own advantages and disadvantages. For example, hierarchical strategy can effectively reduce the

DOI: 10.54254/2753-8818/2026.33183

complexity of tasks and enhance the interpret ability. Offline reinforcement learning can effectively reduce costs, but at the same time, the effect will be discounted with the results. It is mainly applicable to the situation of limited environmental data.

3. Application of a large language model based on reinforcement learning

From the rough introduction of various optimization methods, and can see that people have a relatively mature system in the research of large language models based on reinforcement learning. Next, I will more specifically introduce how reinforcement learning is applied to large language models through four different areas: information retrieval and question answering, text semantic organization and generation, recommendation systems, and graph privacy protection.

3.1. Information retrieval and Q&A

Information retrieval and question answering are widely used fields of large language model training based on reinforcement learning. All fields of society are developing rapidly, and much new knowledge and information have been born in this process. In addition, there is more and more integration in various fields now. When people do not know the relevant information in their unknown fields, they need to use a large language model, an efficient and large-scale information retrieval tool, to achieve information retrieval and question answering. In the field of information retrieval and question answering, the traditional Retrieval-augmented Generation Method(RAG) was generally used in the past. This method has the common problems of the large language model: poor interactivity and weak generalization ability, which make it difficult to land in the reasoning scene. However, this problem has been effectively solved in search-r1 [3]. On the premise of following the core paradigm and classical algorithm of Reinforcement Learning(RL) optimized LLM, it makes key innovations in the scenarios of retrieval, enhanced reasoning and tool use. Search-r1 applies an RL system to the interactive reasoning between LLM and the search engine for the first time. Aiming at the non-differentiability of search operation, it uses trajectory sampling and a token mask to realize the alternate operation of reasoning and multiple rounds of retrieval. Compared with the traditional rag method in the mode of single round retrieval, it will reduce the generation of a lot of irrelevant information. In addition, RAG - Block Guidance Policy Optimization(RAG-BGPO) also solves such problems as low rewriting efficiency and large search space of traditional RAG queries [4]. It also proposes a block-level state action definition on the basis of the traditional RL-LLM, breaking through the binary opposition between token level and sequence level, making the large language model have a better RL granularity as a choice in the face of complex tasks such as multiple rounds of dialogue and tool interaction. These two methods are innovative on the basis of traditional rag, and solve and optimize the pain points of traditional rag in two aspects of the interaction mode and the training mechanism.

3.2. Text semantic organization and generation

The semantic organization and generation of text is the most intuitive evaluation standard to evaluate whether a large language model has excellent reasoning ability. The traditional large language model relies on human annotations to make the large language model benefit human preferences in the process of reinforcement learning, which can be well understood in the RLHF introduced earlier, but the existing method is difficult to make LLM have the ability of independent development reasoning, and it also has a great cost demand for a large number of human annotations

DOI: 10.54254/2753-8818/2026.33183

that need to be given. Because of this, deepseek-R1 came into being [5]. Deepseek-R1 is the product of a pure RL-driven large language model for autonomous reasoning. It breaks the disadvantage of the traditional large language model that needs to rely on a large amount of human annotation data, and effectively solves the problem of high data acquisition cost. At the same time, it skips the traditional process of supervision and fine-tuning, directly takes the pre-training model as the starting point, and advocates rule-based rewards. At the technical level, Group Relative Policy Optimization(GRPO) replaces PPO, reducing resource consumption. In addition to deepseek-R1, which is a pure method of discarding RLHF, some scholars have also studied Direct Preference Optimization(DPO) [6]. The advantage of this method is that it is a lightweight optimization scheme for its human intention. It simplifies the reward mechanism for its human preferences, simplifies the process from monitoring fine-tuning to PPO fine-tuning, and changes to direct optimization, which greatly reduces the cost and has strong stability. It is an excellent, reality-based optimization scheme.

3.3. Recommendation system

In addition to the above two categories, the recommendation system of large language model is also changing from the traditional fixed linguistic preprocessing to data-driven autonomous optimization. A data-centric recommendation agent is one of the new areas of research on this phenomenon [7]. It transforms the fixed template log text into an optimizable module, constructs the training framework of two reinforcement learning stages through GRPO, and takes the accuracy of recommendation as a reward signal to realize the end-to-end optimization of the large language model recommendation system. This research is aimed at the innovation of language. In view of the performance bottleneck of the traditional ID based recommendation system in the scenario of cold start and sparse data, Recommendation Instruction Tuning Language Model (RecLM) also solves the pain point of the traditional recommendation system by effectively combining LLM and Collaborative Filtering (CF) [8]. It conducts customized optimization for LLM, and enables LLM to learn and understand CF, structured interaction and personalization in the process of optimization in terms of task adaptability, representation fusion degree and deviation correction. Using Lora for fine-tuning operation, the lightweight adaptation of the recommendation task is realized on the premise of retaining the general knowledge of LLM, and the recommendation system of LLM is optimized in terms of achievement and efficiency.

3.4. Figure privacy protection

With the development of the Internet, map privacy protection has become a field of increasing public concern. One privacy protection scheme based on deep reinforcement learning uses Open Digital Rights Language (ODRL), a deep reinforcement learning framework. It is mainly to solve the problem of Community Membership Hiding (CMH) [9]. CMH is a key application scenario in graph privacy protection, which is mainly used to prevent privacy from leaking through the graph structure. The deep reinforcement learning framework uses PPO as the main reinforcement learning algorithm, combined with the Generalized Advantage Estimation (GAE), and adopts the decomposition strategy. The core idea of the research and application is to construct a graph named ERD à s – R é NYI for the injection proxy node, connect it with the target node, and realize the edge modification function. Another research on the LLM-guided Dynamic-UMAP (LG-DUMAP) framework is to improve the pain point of machine graph learning in the traditional federated learning scenario [10]. Traditional graph machine learning has some disadvantages, such as limited

DOI: 10.54254/2753-8818/2026.33183

data privacy and sparse graph structure, which are not conducive to graph privacy protection. However, the LG-DUMAP framework integrates the structure embedding of a graph neural network(GNN) and the text embedding generated by a large language model through multimodal embedding, and obtains the node representation with both structural features and semantic information. At the same time, the flow pattern learning of Dynamic-UMAP can also improve the identification of downstream tasks by allowing nodes to gather in a low dimensional space.

4. Existing limitations and future prospects

At present, the optimization method of big prediction models based on reinforcement learning has penetrated into all fields of society, and the research and development in many fields are the beneficiaries of the integration of reinforcement learning and big language models. However, from many studies, some traditional methods are indeed being eliminated with the development of the times, because they can no longer adapt to the current order of magnitude and data cost. At the same time, the traditional large language model relies on the existing fixed data for learning and optimization, and still lacks the ability of independent update and optimize. These are the problems that have puzzled many researchers who are engaged in reinforcement learning and large language model research in recent years. More and more innovative research has been put forward in the past two years, including the four fields mentioned above. These methods can solve the problems of data cost and lack of interpret ability in each neighborhood, and most of the above methods can improve the data effect to a certain extent compared with the original traditional methods in the final evaluation effect, as stated in the previous literature. The optimization of large language models based on reinforcement learning is the inevitable trend of future development. Whether it is based on actual needs or the number of scholars and researchers participating in it in recent years is increasing, which tells us that the optimization of large language models based on reinforcement learning has a good development prospect [11]

5. Conclusion

At present, the research on large language models based on reinforcement learning has a good application and development in many fields. Different fields combine different optimization methods with their own application, and subtly make use of the efficient processing ability of large language models and the combination of reinforcement learning to realize the further development of various industries. The four areas mentioned above are currently of great concern. From their application and development prospects, it can be seen that the large language model based on reinforcement learning is indeed an excellent choice for quantitative and high-quality development in many areas. Currently, large language models based on reinforcement learning still face practical problems such as high data costs, insufficient domain expertise in some areas, and significant black box effects. This indicates that related research still needs further improvement in areas such as data construction, domain adaptation, and interpret ability to better support its practical application in different fields.

References

[1] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.
[2] Schulman, J., et al. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347.

DOI: 10.54254/2753-8818/2026.33183

[3] Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., ... & Han, J. (2025). Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv: 2503.09516.
[4] Mei, W., Xu, G., Zhang, Z., et al. (n.d.). Learning to think in blocks: A prior-guided reinforcement learning framework for RAG.
[5] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv: 2501.12948.
[6] Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36, 53728–53741.
[7] Shi, Y., et al. (2026). From Logs to Language: Learning Optimal Verbalization for LLM-Based Recommendation at Industry Scale. arXiv preprint arXiv: 2602.20558.
[8] Jiang, Y., Yang, Y., Xia, L., et al. (2025). RecLM: Recommendation instruction tuning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15443–15459).
[9] Loi, D., Silvestri, M., Silvestri, F., & Tolomei, G. (2025). Evading Overlapping Community Detection via Proxy Node Injection. arXiv preprint arXiv: 2509.21211.
[10] Puppala, S., Hossain, I., Alam, M. J., Ahad, T., & Talukder, S. (2025). LLM-Guided Dynamic-UMAP for Personalized Federated Graph Learning. arXiv preprint arXiv: 2511.09438.
[11] Alhosani, K., & Alhashmi, S. M. (2024). Opportunities, challenges, and benefits of AI innovation in government services: A review. Discover Artificial Intelligence, 4(1), 18.