Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.20261418

Progress and Analysis of Optimization for Large Language Models Based on Reinforcement Learning

Fengrui Tian*

School of Information and Communication Engineering, Hainan University, Haikou, China

**Corresponding author: Fengrui Tian.*

Abstract

Large language models (LLMs) are one of the current research focuses in society and have extensive applications in various fields. However, when facing complex tasks such as mathematical reasoning at present, it will exhibit problems such as weak generalization ability. Reinforcement Learning (RL) can effectively optimize these issues through reward-guided strategies, thus becoming the core technical path for enhancing the reasoning capabilities of large language models. This article systematically reviews and analyzes the boundaries of how reinforcement learning optimizes the reasoning capabilities of large language models under different resource and scale constraints, as well as the adaptability of reinforcement learning's optimization of large language models in different scenarios. The analysis shows that reinforcement learning algorithms are beneficial for aligning model reasoning and improving the reasoning chain of the model; based on reinforcement learning, strategies such as data minimization training and small model optimization can enhance the resource utilization efficiency of the model during training; however, the improvement of reinforcement learning algorithms on large language models still has problems such as difficulties in expanding the reasoning boundaries.

Keywords

large language model, reinforcement learning, reasoning ability

1. Introduction

Large language models (LLMs) are a type of advanced artificial intelligence model that is trained on vast amounts of text and is capable of deeply understanding, processing and generating coherent natural language. The reasoning ability of LLMs refers to the model's capability of conducting logical reasoning, mathematical calculations, and causal analysis based on the given context information. In recent years, depending on the large-scale pre-training frameworks and powerful self-learning capabilities, LLMs have made remarkable progress in natural language processing tasks such as text generation, and have been widely applied in scenarios such as intelligent education, network services, and intelligent medical assistance. However, when it comes to complex tasks such as mathematical reasoning and logical planning, the model still shows shortcomings such as weak generalization ability, contradictory reasoning steps, and broken logical chains.

Reinforcement Learning (RL) is an optimization technique that optimizes models by actively interacting with the external environment and obtaining reward feedback. It is naturally well-suited to the demand of LLMs for enhancing their reasoning capabilities. Unlike conventional supervised learning, RL does not require a large amount of manually labeled data. The agent can continuously adjust and optimize its strategy by continuously interacting with the environment and receiving reward feedback. The reward feedback is calculated based on indicators such as the correctness of the results and the rationality of the process. In this way, the model can autonomously explore the possibility of reasoning strategies. This continuous optimization in the environment has made RL the core technical approach for addressing the bottleneck of reasoning capabilities faced by LLMs.

This paper systematically reviews the current relevant research results scattered across various fields, constructs a clear and complete analytical framework, clarifies the internal logic of RL in enhancing the reasoning ability of LLMs, summarizes and analyzes the existing optimization methods and practical paths, and establishes a complete and coherent logical analysis chain. The aim is to provide valuable references to practical ideas and directions for optimizing the reasoning performance of LLMs in more practical applications.

2. The Direction of RL in Optimizing the Reasoning Ability of LLMs

2.1 Mainstream RL Algorithms Selection

Among the existing studies, Group Relative Policy Optimization (GRPO) is the mainstream algorithm used. Its core advantage lies in the fact that it does not require a value network, has high sample efficiency, and is suitable for sparse reward scenarios in LLMs' reasoning. For instance, the ARTIST framework [1], the E2H Reasoner method [2], and DeepSeek-R1 [3] all implement reasoning optimization based on GRPO and outperform traditional algorithms such as PPO in tasks like mathematical reasoning and multi-hop question answering.

The Proximal Policy Optimization (PPO) algorithm achieves an optimal balance among training stability, sample efficiency, and computational complexity by constraining the difference between the old and new policies through a clipping mechanism. PPO is particularly suitable for scenarios involving discrete or continuous action spaces, such as the RLPF framework, which uses this algorithm to optimize the personalized inference of user summaries [4].

REINFORCE is a classic Monte Carlo policy gradient reinforcement learning algorithm, belonging to the turn-based policy optimization method. It directly updates the policy parameters by accumulating the rewards of the complete trajectory. Due to its simple implementation, it is used for lightweight tasks, such as small model inference optimization [5].

2.2 Key Innovation Points

In complex reasoning problems, high-quality reasoning data is relatively scarce. For this issue, the research focuses on strategies for minimizing and reusing data. On the one hand, the optimization ability of small-sample RL training is remarkable. For instance, 1-shot RLVR only requires one mathematical reasoning sample for training, and MATH500 pass@1 has increased from 36.0% to 73.6%, approaching the performance of a 1.2k sample set [6], On the other hand, research has shown that when using the data reuse strategy, when the reuse factor τ less than or 25, reusing high-quality data performs comparably to using a large amount of unique data, and the final performance is determined by the total number of optimization steps [7].

Based on this, the dynamic difficulty adaptation strategy has further optimized the training effect of the model. For instance, the E2H Reasoner method proposed a dynamic difficulty scheduling strategy from easy to difficult. By balancing the task distribution through cosine scheduling and Gaussian scheduling, the model gradually optimizes the strategy, resulting in an increase of 32.9% in the accuracy of small model tasks [2]; while ALP dynamically adjusts the length penalty according to the task difficulty, allocating a larger length penalty for simple tasks and relaxing the constraints for complex tasks [8].

Meanwhile, multi-stage and cross-modal training is also one of the key innovation points. For instance, DeepSeek-R1 has designed a two-stage training framework. The training in the first stage enhances the model's autonomous reasoning ability through pure RL algorithms, and in the second stage, it achieves integration with

human preferences by combining SFT and secondary RL, achieving a pass@1 rate of 77.9% on the AIME 2024 dataset [3]; FinLMM-RL integrates text and image information and designs a multimodal reward to expand the application of RL in financial scenarios [8].

3. The Adaptability of Different RL Algorithms in Different Scenarios

3.1 Adaptability of Task Types

Firstly, tasks such as mathematical reasoning and code writing have definite answers with clear judgment methods, which are the core scenarios of RL optimization. DeepSeek-R1 achieved a pass@1 rate of 77.9% on the AIME 2024 dataset (and even 86.7% under self-consistency decoding), surpassing the average level of humans [3]; moreover, the research proves that 32B parameters are the balance point for the model to pursue the efficiency and performance of mathematical reasoning problems, and the 72B model needs more training steps due to efficiency saturation to surpass [7].

In multiple tasks such as question answering and retrieval reasoning, the ReSearch framework embeds the search operation into the reasoning chain, granting the model the autonomy to search during the reasoning process. End-to-end training on the GRPO, it helps the model understand when to search and how to do so. On the Bamboogle dataset, the EM reaches 56.8%, which is 25.6% higher than the baseline standard [9]. Meanwhile, RAG-Star combines MCTS and retrieval verification, which can enhance the integration of internal and external knowledge within the model, helping it further improve its ability to integrate data from different sources and enhance the reliability of complex reasoning [8].

In the field of security and personalized reasoning, CI-RL, for scenarios where information is prone to disclosure, through the CI-CoT prompt strategy and RL optimization, can significantly reduce the rate of privacy information leakage compared to the evaluation benchmark of authoritative privacy protection tools [8], for the direction of personalized reasoning [10], RLPF mainly realizes personalized summaries for users, with the subsequent task prediction performance as the reward, and can achieve a 74% reduction in context compression while increasing the cross-domain transfer accuracy by 15.68% [4].

3.2 Model Scale Adaptability

Regarding the impact of parameter size on model performance, existing research has identified a notable threshold. When the model reaches a parameter size of over 32B, its utilization efficiency of training data significantly improves. Experiments have shown that a 72B model can achieve a performance level similar to that of a 1.5B model with only about 10k samples [7]. This indicates that when aiming for the same effect, large models actually require fewer total data volume. However, this advantage comes at a cost. After parameters exceed 32B, the performance gains from further expanding the scale gradually slow down, and there may also be cross-domain negative transfer problems, that is, optimizations in certain new tasks may weaken the model's performance in general scenarios.

In contrast, small models with a size below 10B have relatively low data utilization efficiency, but they have the advantage of low training costs and fast convergence. Some studies have demonstrated that with 7,000 samples and approximately $42 in cost, a 1.5B model was able to increase its accuracy on the AMC23 dataset from 63% to 80%, even surpassing the level of o1-preview at that time. However, these models are not very stable in practical use. Once the training steps exceed a certain number, performance may decline, and it is necessary to balance the data difficulty and length when using them [7].

4. The Fundamental Limitations of RL Optimization

4.1 For the Breakthrough in Optimizing the Inference Boundaries of RL

Whether reinforcement learning can truly break through the upper limit of the reasoning ability of LLMs has not reached a unified conclusion in the academic community. Some researchers have pointed out that the main role of RL is actually to improve sampling efficiency, and it is difficult for the model to generate reasoning paths that it originally does not possess. Relevant experiments also show that as long as the sampling times of the base model are sufficient, its performance can even be better than the model optimized through

reinforcement learning. This also to some extent indicates that the correct reasoning methods found by RL essentially still come from the generation distribution of the base model itself [7, 11].

Some scholars hold different opinions. Research has shown that after training the model using mixed-domain data, the model can acquire new reasoning skills in new domains with lower similarity. For instance, a model trained on the GURU dataset can expand the reasoning capabilities of the original base model on tasks such as the Zebra Puzzle [12].

This article holds that the core significance of reinforcement learning lies more in activating the existing reasoning paths of the model and improving the reasoning efficiency. However, after integrating methods such as cross-domain data fusion, the model can also, through the optimized reasoning method, solve problems that were previously impossible to handle in low-similarity new domains. The final effect is often closely related to the specific training strategy and the characteristics of the task itself.

4.2 The Current Key Challenges

RL is based on the reward feedback provided by the environment, and it inherently has strong task specificity. Its optimization effect on mathematical reasoning tasks is difficult to transfer to domains with significant differences such as logic and simulation [7, 12]. Currently, only task types such as mathematics and programming, which frequently appear during pre-training, can achieve rapid improvement through cross-domain reinforcement learning, while tasks like logic and simulation often require specialized training within the corresponding domain to obtain practical and effective performance gains. Moreover, the current benchmarks used to evaluate the reasoning ability of models still have many flaws, such as data contamination and misaligned reward signals, making it difficult for the evaluation results to truly reflect the actual reasoning level of the model. Although dynamic evaluation benchmarks can alleviate the impact of data contamination by capturing new problems in real time, there is still a lot of room for improvement in the design of evaluation indicators during the reasoning process.

5. Conclusions

The core of RL optimizing the inference of LLMs lies in amplifying the existing capabilities of the model itself. The GRPO algorithm is the main means to achieve this goal, while data optimization is to better adapt to resource constraints. On this basis, the model's parameters can reach a balance of benefits around 32B. The advantage of large models in sample utilization, which is complementary to the low-cost gain characteristic of small models, enables different-sized models to have their own value in practical applications. It should be noted that RL does not create completely new inference capabilities that the model does not originally have; rather, it is more like amplifying the potential performance of the base model. Current research still has issues with unbalanced optimization effects, and problems such as weak cross-domain transfer ability and evaluation data contamination need to be addressed simultaneously.

For the development of future large language models, people need to integrate dynamic adaptive reward functions to alleviate various biases, build multi-type integrated inference datasets and domain-general reward mechanisms to enhance the model's generalization ability, and explore the paths of few-shot cross-domain reinforcement learning and small model distillation transfer to further lower the resource threshold for technology implementation.

References

[1] Singh, J., Magazine, R., Pandya, Y., & Nambi, A. (2025). Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*.

[2] Narvekar, S., & Stone, P. (2018). Learning curriculum policies for reinforcement learning. *arXiv preprint arXiv:1812.00285*.

[3] 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*.

[4] Wu, J., Ning, L., Liu, L., Lee, H., Wu, N., Wang, C., Prakash, S., O'Banion, S., Green, B., & Xie, J. (2025). RLPF: Reinforcement Learning from Prediction Feedback for User Summarization with LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(24), 25488-25496.

[5] Dang, Q. A., & Ngo, C. (2025). Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't. *arXiv preprint arXiv:2503.16219*.

[6] Wang, Y., Yang, Q., Zeng, Z., Ren, L., Liu, L., Peng, B., ... & Shen, Y. (2025). Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

[7] Tan, Z., Geng, H., Yu, X., Zhang, M., Wan, G., Zhou, Y., ... & Bai, L. (2025). Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*.

[8] Pan, P. C., Liang, Y., & Lin, S. (2026). Reward Modeling for Reinforcement Learning-Based LLM Reasoning: Design, Challenges, and Evaluation. *arXiv preprint arXiv:2602.09305*.

[9] Chen, M., Sun, L., Li, T., Sun, H., Zhou, Y., Zhu, C., ... & Chen, W. (2025). Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

[10] Lan, G., Inan, H. A., Abdelnabi, S., Kulkarni, J., Wutschitz, L., Shokri, R., ... & Sim, R. (2025). Contextual integrity in LLMs via reasoning and reinforcement learning. *arXiv preprint arXiv:2506.04245*.

[11] Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., & Huang, G. (2025). Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?. *arXiv preprint arXiv:2504.13837*.

[12] Cheng, Z., Hao, S., Liu, T., Zhou, F., Xie, Y., Yao, F., ... & Hu, Z. (2025). Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *arXiv preprint arXiv:2506.14965*.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access