Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33177

Reinforcement Learning Combined with Tree Structure Search Enhances the Reasoning Ability of Large Language Models

Junhao Xue

School of Computer Science and Engineering, Central South University, Changsha, China
8208230210@csu.edu.cn

Abstract. The combination of tree structure search and reinforcement learning reasoning process is a reinforcement learning method by introducing intermediate steps into the reasoning tree to achieve no additional introduction of reward model conditions. This method reduces the risk of credit distribution and reward deception to a certain extent. This paper systematically reviews and analyzes the research direction of strong chemistry combined with tree structure search to improve the reasoning ability of large language models. After conducting relevant research, this paper focuses on two issues: whether the tree search mechanism actually changes the reasoning structure of the language model. This method is difficult at the level of computational cost and scalability. Comprehensive research shows that the core value of the combination of tree structure search and reinforcement learning is to transform the reasoning process from linear generation to a more complex structured decision problem, which gives more intensive process signals for reinforcement learning, but the performance improvement still depends on the basic model's ability and search budget control.

Keywords: Online policy tree search, reinforcement learning, large language model, reasoning ability

1. Introduction

The powerful ability of the large language model has been reflected in various complex reasoning tasks, including the direction of mathematics, programming and autonomous agents [1]. And reinforcement learning has become an important way to improve the ability of large language models by virtue of the reward feedback optimization mechanism [2].

In order to avoid the limitation of single path chain sampling, many methods now take the independent sampling of multiple tracks and obtain the reward signal based on the final answer [3]. However, although tree structure search has achieved remarkable results in the field of Go, it has developed slowly in the field of large language models. Existing related work focuses on two directions: first, combining with external rewards, using tree search to improve the performance of the model reasoning stage [4]; The second is to model agent interaction with the help of tree structure to provide process supervision [5]. So far, the potential of integrating tree search into

© 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: 10.54254/2753-8818/2026.33177

online strategy intensive learning and training to improve the reasoning ability of large language models has not been fully explored.

It can be seen that the existing research has not systematically clarified the role of tree structure search in the process of reinforcement learning training, that is, whether it can exist as a part of the training framework rather than only as an external auxiliary tool. Further, the direct integration of tree search into the online strategy reinforcement learning framework enables the model to obtain process level reward signals through structured exploration in the training phase, which may provide a new implementation path for alleviating the difficulty of credit allocation and improving the efficiency of exploration.

Based on the above background, this paper will focus on the research progress of the combination of reinforcement learning and tree structure search, aiming to clarify the improvement of pure reinforcement learning by the fusion of tree structure, and the enhancement of the reasoning ability of large language model by the new paradigm of the combination of the two; At the same time, it explores the essential reasons for the improvement of the paradigm and analyzes the existing problems at this stage, so as to explore the possibility of continuing to improve the large language model in the future.

2. Basic information

2.1. Research context of tree search enhanced reasoning

As a method of optimizing decision-making strategy based on reward feedback, reinforcement learning is very suitable for training and improving the reasoning ability of language models. In the early stage of reinforcement learning, RLHF is mainly used to improve reasoning ability by aligning and optimizing the final output. However, this kind of method lacks effective improvement and modeling of the reasoning intermediate process. To this end, the researchers introduced the tree search method, represented by MCTs, to evaluate and select the intermediate steps by constructing the path search space [6]. This improvement not only improves the reasoning performance but also generates high-quality process monitoring data for model training or combined with the DPO optimization step level preferences [7]. At the same time, aiming at the problems of search efficiency and interpretability, people further put forward the improvement methods for the structure, such as introducing the comparative reward model and optimizing the node selection strategy [8]. Starting from the existing strategies, the current research gradually explores the deep integration of tree search and model training. On the one hand, the strategy guides tree search to realize dynamic search decision learning and reduce the dependence on heuristic rules; On the other hand, using the AlphaZero paradigm for reference, people combine value function learning to achieve iterative optimization of reasoning and training [9]. In addition, in combination with reinforcement learning, tree search is also used to generate high-quality reasoning trajectories. As enhanced data, the trajectory can improve the ability of the model to deal with complex reasoning tasks and promote the transformation from simple "result optimization" to "process modeling".

2.2. Limitations of existing research

Overall, the existing research has undoubtedly made rich progress in using tree structure search to improve the reasoning performance of large language models, including process search path optimization, supervision data generation, reward model design and so on. However, these methods often only rely on tree search to enhance the reasoning stage or use it as an offline data generation

DOI: 10.54254/2753-8818/2026.33177

mechanism, and do not fundamentally solve the deep integration problem of tree search and online RL reinforcement training. Therefore, how to directly incorporate the structured search mechanism into the reinforcement learning training framework and enable the model to obtain more intensive and effective feedback through structured exploration in the training process is still the most noteworthy issue in the current research.

3. Method and mechanism analysis

3.1. On-policy tree RL

This kind of method attempts to integrate the tree structure search directly into the reinforcement learning training process, so that the model can obtain the process level reward signal through structured exploration in the training stage, rather than relying only on the termination result supervision. In the treerl framework proposed by Hou et al., new nodes are generated iteratively from the first n tokens with the highest entropy in the existing search tree to realize the expansion of the search tree. In the training process, each reasoning step is assigned a credit score based on the dominance value, and the process signal is weighted by the global dominance value and the local dominance value. These advantage values are directly obtained through the online strategy tree without relying on other reward models. This architecture is based on the trained qwen-2.5-14b-sft model and tested on the omni-math-500 data set with the pass rate as the target index. The results show that under different generation costs, the performance of eptree is always better than Monte Carlo tree search, and is 3% higher than multi-chain sampling [10].

3.2. MCTS+RL(self-improvement)

This method adopts the "Learning+search" fusion paradigm represented by AlphaGo/AlphaZero. The core logic is to use the value/strategy model obtained from learning to guide the exploration and utilization of tree search, and then optimize the model through the results of tree search, so as to realize the mutual promotion of learning and search. After being proposed, this method has undergone many improvements to adapt to the multi-step reasoning environment of LLM. Many experiments have proved that the combination of Monte Carlo tree search and enhanced self-training can significantly improve the reasoning ability of large language model. In the sc-mcts framework proposed by Gao et al., multirm is constructed to normalize the context of each reward model by fine-grained prior statistical characteristics of reward experience distribution. Based on the llama-3.1-70b model, the average accuracy of the reasoning method in the building block world dataset is 17.4% higher than that of the O1 mini model, which verifies its superiority [8].

3.3. Test-time tree search

This method does not modify the model, but applies tree search enhancement in reasoning. As an innovative reasoning mode, it lays the foundation for the subsequent method of combining tree search and reinforcement learning. In Yao's tree of thoughts, the tot framework proposed by Yao as a general solution to the language model has changed the previous basic autoregressive text generation mode. By maintaining a thinking tree, the "thinking" of each node is a coherent text sequence, which is used as an intermediate step for problem solving. In this reasoning method and intermediate steps, a thinking generator and a state evaluator are designed, and BFS and DFS are initially used as search algorithms. Compared with the traditional baseline methods such as input-output promotion and thinking chain prompt, the success rate of this reasoning method on the 24-

DOI: 10.54254/2753-8818/2026.33177

point game data set increased from 9% to 74%; Increased from less than 16% to 60% in mini crossword puzzles [4].

3.4. Common datasets

Table 1. Commonly used test sets for reasoning ability of large language models

Dataset	Data scale	Depth of Reasoning	Dataset problem
MATH-500	500 carefully selected challenging math problems	5-15steps	The annotation process was noisy, and some answers were not standardized.
GSM8K	8.5k Elementary School Word Problems	2-5steps	Small branch space, strong template deviation
OlympiadBench	Olympiad level difficulty	10-30steps	Rewards are sparse, making it difficult to assess the correctness of reasoning paths.
LivecodeBench	Real-time updated code questions		Reward hacking

Table 1 shows the common different data sets currently used to test the reasoning ability of the model, including tasks with different difficulties, and the reasoning steps required between different data sets are quite different. Therefore, the above data sets can effectively verify the effectiveness and robustness of the reasoning ability of the model.

4. Existing problems and possible solutions

4.1. Tree search may not have truly changed the policy space

While the current "tree search + deep learning" approach introduces branch exploration, the underlying language model remains autoregressive：

$$\pi\left(a_t \mid S_t\right) \tag{1}$$

Equation (1) shows that the policy function is essentially a linear sequence decision. Therefore, the question that needs to be addressed is whether tree search is merely an external construct, rather than altering the internal linear generation of the model.

4.2. Credit allocation has not been truly alleviated

Tree structure search mostly uses the value of intermediate nodes to evaluate and improve credit allocation. However, in the LLM scenario, the node value often comes from its own estimation or from the additional reward model, which makes the model more inclined to expand the preference path and form the result of reward deception. That is, what the model learned is "the strategy of improving the intermediate score", not improving the reasoning quality. In the tree structure search, the risk is further expanded because multiple backtracking updates of the tree structure will strengthen the model's local error estimation and spread the value of the error along the tree.

DOI: 10.54254/2753-8818/2026.33177

4.3. Search complexity and scalability issues

The theoretical complexity of tree search is approximately:

$$O\left(B^d\right) \tag{2}$$

As shown in formula (2), B is the number of branches and d is the depth; in language models, d can reach hundreds, and its action space far exceeds that of chess games. Therefore, the performance improvement obtained in the current experiment may essentially be dependent on computing power. Under a fixed budget, this method may no longer have an advantage.

4.4. Solution

To solve the above problems, future research can promote the deep integration of tree search and reinforcement learning in language model reasoning from multiple directions. First, at the level of strategy space, the explicit hierarchical decision structure should be introduced to integrate the original autoregressive strategy.

It is extended to a hierarchical strategy with "high-level planning - low-level generation", so that the model can explicitly model the reasoning path structure internally, instead of relying only on the external tree search for a posteriori correction, so as to truly change the strategy expression ability. Secondly, on the issue of credit allocation, it is necessary to introduce verification signals closer to the real environment, such as objective rewards based on the feedback of executable programs, symbol solvers or external tools, so as to reduce the dependence on learning value functions or reward models and alleviate the problems of reward deception and wrong value propagation. At the same time, people can combine uncertainty estimation or confidence calibration mechanism to constrain the value of nodes in the tree, so as to avoid local misjudgments being amplified in multiple backtracking. Finally, in terms of search complexity and scalability, people should systematically evaluate the performance of the method under the constraint of fair computing power, introduce budget aware search strategy and adaptive branch control mechanism, for example, dynamically determine the expansion depth and width through the policy network, or combine pruning and value guidance to reduce invalid exploration, so as to achieve the balance between performance and efficiency under limited computing resources.

5. Conclusion

In recent years, the combination of reinforcement learning and tree structure search has become a hot research topic. Different from the traditional single path chained sampling analysis, the combined method transforms the reasoning process into a structured decision problem by fusing the multi-branch reasoning space created by tree structure search, and provides a more intensive intermediate signal and diversified exploration path for the large language model. This paradigm has achieved some performance improvements in mathematical reasoning, code generation and other tasks, and has largely solved the problems of credit allocation and sparse reward signals.

From the perspective of methodology, this direction mostly attempts to emphasize structured exploration, making it an important step in strategy optimization. As an enhancement tool in the reasoning stage, or as a generation mechanism of process supervision, or even online integration of the strategy tree into RL, this direction provides a new research perspective for reinforcement

DOI: 10.54254/2753-8818/2026.33177

learning. However, through the analysis, it is found that most of the current methods have not fundamentally changed the linear structure of the language model autoregressive decision-making, and tree search is more about the role of sampling redistribution and value estimation enhancement, rather than the reconstruction of the expected reasoning space.

Therefore, future research directions should focus on three key points: first, embedding structured decisions into the model to avoid only "external search"; Second, introduce more environmentally binding authentic feedback quotes to reduce the risk of reward fraud; Finally, the paper attempts to make a systematic comparative study under strict budget control to clarify the actual revenue boundary of structured exploration.

In general, although the integration of reinforcement learning and tree structure search has achieved many results, it is still in the exploratory stage. However, its representative idea of structural decision optimization provides significant inspiration for the subsequent improvement of the reasoning ability of large language models. Whether this direction can become the mainstream paradigm of the subsequent reasoning model still needs a deeper breakthrough in the mechanism level, and can be effectively verified in the future large-scale real tasks.

References

[1] Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., ... & Misra, V. (2022). Solving quantitative reasoning problems with language models. Advances in neural information processing systems, 35, 3843-3857.

[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

[3] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., ... & Cobbe, K. (2023, May). Let's verify step by step. In The twelfth international conference on learning representations.

[4] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36, 11809-11822.

[5] Zhang, D., Zhoubian, S., Hu, Z., Yue, Y., Dong, Y., & Tang, J. (2024). Rest-mcts*: Llm self-training via process reward guided tree search. Advances in Neural Information Processing Systems, 37, 64735-64772.

[6] Li, S., Dong, S., Luan, K., Di, X., & Ding, C. (2025). Enhancing reasoning through process supervision with monte carlo tree search. arXiv preprint arXiv: 2501.01478.

[7] Xie, Y., Goyal, A., Zheng, W., Kan, M. Y., Lillicrap, T. P., Kawaguchi, K., & Shieh, M. (2024). Monte carlo tree search boosts reasoning via iterative preference learning. arXiv preprint arXiv: 2405.00451.

[8] Gao, Z., Niu, B., He, X., Xu, H., Liu, H., Liu, A., ... & Wen, L. (2024). Interpretable contrastive monte carlo tree search reasoning. arXiv preprint arXiv: 2410.01707.

[9] Feng, X., Wan, Z., Wen, M., McAleer, S. M., Wen, Y., Zhang, W., & Wang, J. (2023). Alphazero-like tree-search can guide large language model decoding and training. arXiv preprint arXiv: 2309.17179.

[10] Hou, Z., Hu, Z., Li, Y., Lu, R., Tang, J., & Dong, Y. (2025, July). Treerl: Llm reinforcement learning with on-policy tree search. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12355-12369).