From Verifiable Rewards to Autonomous Evolution: Reinforcement Learning-Driven Large Language Model Reasoning Abilities

Mengbo Song

Software Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi, China

songmengbo2024@outlook.com

Abstract. This paper attempts to outline the evolution of LLMs from classic methods of supervised finetuning models with static human-annotated datasets, to a more dynamic and evolutionary reinforcement learning based autonomous models. Along with the systematic evolution of reasoning capabilities, this is one of the most prominent focal points in the field of AI. This paper attempt to outline the most current advancements in reasoning and reinforcement learning, and classify the occurrences into the applicable areas of criteria, such as: choice of architecture, choice of reward assignments, and choice of evaluation metrics. This paper explore reasoning improvements as a result of self-reflective and exploratory processes within the bounds of the RL with Verifiable Rewards (RLVR) framework. By systematically studying the aforementioned criteria across multiple works and the respective variations in algorithmic efficiency, control of reward signal bias, and performance metrics, this paper want to outline the positive role of reinforcement learning in fostering autonomous self-correction in models and complex thought chain processes. The aim of this paper is to describe the potential autonomous advancements the next generations of large language models may evolve and want to offer some suggestions as a theoretical and a technical framework.

Keywords: Large Language Models, Reinforcement Learning, Verifiable Rewards, Logical Reasoning.

1. Introduction

The reasoning capabilities are obtained through supervised fine-tuning (SFT) in traditional large language models. This paradigm is basically a copy of human thought patterns; it is not only costly, but also has a limit due to the quality and variety of annotated information. The recent studies, embodied by DeepSeek-R1, have started the next phase of reasoning reinforcement based on pure reinforcement learning [1]. These experiments illustrate that models can have higher-level reasoning patterns of reflection, verification, and strategy changes, when they are motivated by verifiable signals of reward. The advent of this slow thinking ability drives LLMs towards substantial logical construction as opposed to the mere probability prediction.

Reinforcement learning (RL) offers a more dynamical optimization model to optimize the reasoning abilities of large language models compared to other supervised fine-tuning techniques that use manually-generated reasoning examples. The essence of it is that it is no longer merely a process of replicating the existing ways of thought, but is instead a process of exploration, correction and selection, with an ongoing feedback of rewards, allowing it to build more adaptive thinking strategies. It is particularly in tasks that can be verified that this mechanism allows the model to self-tune its output process in a circular feedback loop of " out-feedback-re-optimization, " thereby providing an evolutionary ability unique to SFT traditionally. This is why, RL has gradually become a significant technical method of contemporary reasoning-oriented LLM studies.

Nonetheless, even though reinforcement learning elicits model potential, it also reveals the brittle nature of the reward mechanism design, including the so-called Reward Hacking effect, in which models cheat to score highly by repeating actions that do not constitute real reasoning logic [2]. Moreover, the variability of sampling efficiency, noise of reward signal and the constraint of cross-domain generalization are fundamental issues of the existing Reinforcement Learning with Verifiable Rewards (RLVR) paradigm [3, 4]. The paper at hand will attempt to review in a systematic way the

recent directions in which RL can be improved to be able to reason with models in terms of multi-dimensional comparison, and whether autonomous evolution is a possibility.

2. Differences in algorithmic architectures across experiments

In RL-based reasoning research, different algorithms impact the model's scalability with CoT training. While traditional PPO algorithms are stable, they are highly memory intensive due to the additional value network when dealing with long output sequences. To alleviate this limitation, experiments with DeepSeek-R1 conducted an initial trial with the GRPO algorithm and successfully removed the critic network and instead adjusted relative advantages within a group for a given prompt, leading to improved efficiency [1].

Compared to this network's large-scale universal design, the Logic-RL experiment for structured logical puzzles utilized the REINFORCE++ algorithm, and due to the addition of KL divergence penalties and improved gradient updates, it was able to converge more quickly than PPO while maintaining lightweight design [5]. Furthermore, to resolve some reasoning knowledge gaps, the ReSearch experiment integrated a search process within the RL chain and trained the model to make decisions about when to perform external search retrieval, thereby transforming the logic from purely internal reasoning to also include external tools [6].

3. Design logic and bias control of reward mechanisms

As the evaluation criterion for RL, the core contradiction in reward mechanism design lies in how to provide feedback that is both accurate and guiding. Currently, rule-based verifiable rewards (RLVR) dominate in mathematics and programming tasks, providing hard constraints by checking the final correctness of answers (outcome rewards) [7]. However, single outcome feedback often leads models to ignore reasoning logic for the sake of scoring, producing so-called "reward hallucinations." Su et al. (2025) further quantified this issue in a cross-domain RLVR study: among real-world exam questions, only 60.3% of mathematical problems and 45.4% of multi-domain problems could be strictly verified by rule-based methods, indicating that sole reliance on outcome feedback is insufficient in open-ended scenarios [8].

Reward models, particularly Process Reward Models (PRM), can theoretically provide denser supervisory signals, but in practice, they are easily exploited, leading to serious Reward Hacking issues where models generate many correct but meaningless repetitive steps [2]. Furthermore, relevant boundary analysis research points out that the effectiveness of verifiable rewards highly depends on the internal knowledge boundary of the base model; when a model lacks corresponding prior knowledge, it is difficult to induce correct reasoning paths solely through reward signals [7].

Consequently, researchers have begun introducing format rewards as auxiliary means, forcing models to reason within specific "thinking tags," which significantly reduces logical jumps [5]. Combining the two ensures that result rewards secure the upper limit of reasoning accuracy, while format rewards ensure the lower limit of process logic. To combat reward hacking, techniques such as Clip and Delta have been proposed to provide an optimization objective with an upper bound, preventing models from meaningless reward exploitation [2]. These explorations reflect attempts to build a unified "Reasoning-Aligned Reinforcement Learning" (RARL) framework to systemically solve reward signal design problems [9].

4. Evolution of datasets and experimental evaluation standards

The evaluation of reasoning capabilities is shifting from single-domain to more challenging multi-domain verification. Early research mostly concentrated on mathematical sets like MATH and GSM8K, but recent research such as the GURU corpus has expanded the scope to six major fields, revealing that the transfer efficiency of RL across different domains varies significantly [1]. The SWE-RL experiment directly utilized code snapshots of Pull Requests from GitHub for RL training, allowing models to better adapt to actual software engineering scenarios [10].

To capture the true performance of models in long-chain thinking, evaluation standards have evolved from Pass@1 to CoT-Pass@K, the latter of which excludes samples that might get the correct answer by chance by verifying both the logic of the thought chain and the final answer [7]. Notably, some research indicates that while current RL significantly improves sampling efficiency, it is still largely limited by the base model; therefore, one should focus on whether the model breaks through original capability boundaries at extremely high sampling rates (large K values) rather than just score improvements [3].

5. Future challenges and possible solutions

The reasoning models based on RL are still plagued by three bottlenecks. To begin with, it is hard to evaluate the accuracy of partial answers in areas that do not have objective guidelines, e.g., literature creation. Breakthroughs in the future might be in the creation of a verification system in which more than one model is a referee to the other, which produces feedback via semantic entailment analysis and backward logical traces.

Second, existing modes of reasoning are closed in nature. The further development is toward the direction of what is called Agentic Reasoning, in which models have the capability to summon interpreters to trial and error, can see execution logs and can automatically fix thought chains immediately through external feedback. Lastly, on the topic of "reward hallucinations, " atomic checks of all derivation steps may be investigated in future research with symbolic logic verifiers. It is only then that a model comes to the realization that the faults in logical steps are what cause penalties and hence it will actually be a change of guessing answers to deriving logic. "

Moreover, existing studies in the reinforcement learning (RL) field on enhancing the reasoning capacity of large language models are predominantly based on highly structured and rule-based task domains, including mathematical proofs, code completion, and solving logic puzzles. Nevertheless, various fields vary considerably in terms of the knowledge structure, the form of problems and the verifiability of feedback, so it is challenging to make the policies that are learned as part of RL training transfer easily to less restricted and more complex situations. Models can continue to exhibit serious policy degradation, thought fragmentation, or reward failure, especially in interdisciplinary reasoning, in the real-world of engineering decision making, and interactive task planning.

Further future research must enable the development of a single cross-domain training architecture to advance the transferability of models on three levels: task design, reward modeling and evaluation benchmarks. On the one hand, it can be done by having more hybrid, multi-hop, and contextualized tasks to enhance the model to adapt to various problem structures. Conversely, a complete evaluation set that includes multiple situations like mathematics, programming, reasoning question answering, tool invocation and real-world planning is also required to investigate the strength and transferability of the reasoning strength of the model in a more comprehensive way.

In the meantime, as the current systems of evaluation develop to include Pass@1 into more sophisticated performance criteria such as CoT-Pass@K, the current focus on the correctness of the ultimate response can obscure logical fallacies, pseudo-reasoning and chance hits during the intermediate stages of the model. On the other hand, concentrating on formal integrity of chain-like reasoning can easily falsely assume that lengthy expressions are considered to be reasoning.

Thus, the standards of evaluation in the future should be gradually changed to multi-dimensional joint evaluation. The accuracy and stability of the final result should be investigated on the one hand;

the traceability, consistency and verifiability of the reasoning process in the middle should also be taken into account on the other hand. Moreover, it is important to consider the varying performance of the model at various sampling intensities, task challenges, and domain conditions to prevent the model just gaining shallow benefits under one benchmark or a particular prompt template.

6. Conclusion

Reinforcement learning has successfully evolved LLMs from passive learners into active explorers. Through algorithmic optimization, refinement of reward mechanisms, and cross-domain data, models have made significant progress in logical depth and self-correction. However, this review also reveals internal tensions: while RL improves sampling efficiency and performance, issues like reward hacking and the capability boundaries of base models remain. The paradigm of achieving capability enhancement through verifiable reward feedback has been fully confirmed. Future tasks lie in breaking through non-verifiable domain limits and transforming static reasoning into dynamic decision-making in complex environments. As reasoning verification rules deepen, LLMs will eventually achieve autonomous evolution beyond human reasoning boundaries.

References

[1] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., ... & Tan, Y. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 645(8081), 633-638.

[2] Gao, J., Xu, S., Ye, W., Liu, W., He, C., Fu, W., ... & Wu, Y. (2024). On designing effective rl reward at training time for llm reasoning. arXiv preprint arXiv:2410.15115.

[3] 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.

[4] 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.

[5] Xie, T., Gao, Z., Ren, Q., Luo, H., Hong, Y., Dai, B., ... & Luo, C. (2025). Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2502.14768.

[6] 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.

[7] Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., ... & Yang, M. (2025). Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. arXiv preprint arXiv:2506.14245.

[8] Su, Y., Yu, D., Song, L., Li, J., Mi, H., Tu, Z., ... & Yu, D. (2025). Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. arXiv preprint arXiv:2503.23829.

[9] 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.

[10] Wei, Y., Duchenne, O., Copet, J., Carbonneaux, Q., Zhang, L., Fried, D., ... & Wang, S. I. (2025). Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. arXiv preprint arXiv:2502.18449.