

8 Conclusions and Future Directions

In this paper, we have introduced the fundamental concepts of RL from the perspective of LLM training. We began with basic RL formulations and algorithms, including policy gradients, advantage estimation, importance sampling, and reward modeling, and explained them through LLM training examples. We then discussed recent advances in RL. We further extended the discussion to reasoning models, agentic systems, and multimodal models, showing how RL has evolved from a general policy optimization framework into a key paradigm for enhancing LLM reasoning, interactive decision-making, and multimodal generation capabilities.

While RL has become well established in LLM training and has achieved remarkable results, several promising directions remain for future exploration:

- **RL for pre-training.** Most existing studies focus on applying RL during post-training, while its role in pre-training remains less explored. Extending reward-driven learning to the pre-training stage may provide a way to shape model capabilities earlier in the training process. Recent studies have begun to demonstrate the potential of this direction (Dong et al., 2025), but achieving stable and scalable optimization at pre-training scale remains an important challenge.
- **More efficient RL training.** Recent methods, such as GRPO, have simplified RL training by removing components such as the critic model. However, RL still relies heavily on online sampling, which introduces substantial computational and time costs. Improving sampling efficiency and reducing the overall training cost therefore remain important directions for future research.
- **Predicting the scaling limits of RL.** The final gains from RL depend on many factors, including model capability, training data, reward quality, and optimization settings. In practice, we often need to complete expensive RL runs before knowing the final performance. Developing methods to predict the potential gains and performance limits of RL before full-scale training could significantly improve the efficiency of model development.
- **Continual and self-evolving RL.** Most current RL pipelines optimize a model on a fixed training distribution and stop once training is completed. One more ambitious direction is to enable models to continuously learn from their own interactions, feedback, and accumulated experiences after deployment. This requires new mechanisms for experience selection, memory and skill evolution, and stable policy updating without catastrophic forgetting.

References

Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training, 2025. URL <https://arxiv.org/abs/2506.08007>.