

1 Introduction

RL is an interdisciplinary field, and has its origins in both psychology and optimal control. Over the years, the concept has been further developed and formalized within the realms of artificial intelligence and machine learning. The basic idea is that an agent learns to make decisions by receiving feedback on its actions from the environment. This approach is very general and applies to a wide range of problems, from playing complex games like chess and Go to controlling autonomous vehicles.

A recent breakthrough is the large-scale application of RL in the field of natural language processing (NLP), in particular in the development of LLMs. For example, RL has been widely considered an effective way of aligning pre-trained LLMs with human preferences and values (Christiano et al., 2017). One such method, called reinforcement learning from human feedback (RLHF), forms the basis for the development of advanced LLMs like OpenAI’s GPT-4. More recently, RL has shown great promise in enabling LLMs to perform complex reasoning (OpenAI, 2024; Deepseek, 2025; Team et al., 2025). As a result, interest in RL for LLMs has exploded. Numerous conference papers now discuss RL in LLMs, with even more appearing in the past few months. The research community is highly enthusiastic, and many in the field are anticipating a new turning point where large-scale RL algorithms could further advance AI. This trend is further strengthened by the rise of LLM-based agents, where models must learn from interaction and experience to make effective decisions in dynamic environments. From this perspective, RL is becoming a key technique for building more capable and adaptive intelligent systems. This view closely echoes the “Reward Is Enough” hypothesis proposed by Silver et al. (2021):

“Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment.”

This hypothesis highlights why RL is particularly relevant to the next stage of LLM development: instead of learning only from static supervision, models can acquire increasingly complex capabilities through large-scale experience and rewards.

Historically, RL has played a relatively limited role in mainstream NLP, where supervised learning has long been the dominant paradigm. Although applying RL to LLM training seems a natural choice for machine learning researchers, it introduces many new concepts to the NLP community, which has traditionally been dominated by supervised learning methodologies. As NLP researchers, when we attended presentations on LLMs, we often heard statements such as “use a value function to estimate the expected cumulative reward”, “learn the policy with PPO”, or simply “we need to train a reward model”. These techniques were often presented as if they required little explanation. Yet, for researchers trained primarily in supervised learning, the mathematical formulations of RL, with their many expectations, value functions, and advantages, can be difficult to follow. This gap becomes increasingly important in LLM training, where understanding RL is no longer a specialized topic, but is becoming essential for following and developing modern AI systems.

It is natural to learn RL from standard references, which appears straightforward at first. However, common RL textbooks, such as Sutton & Barto (2018)’s book, are mostly based on classic robotics or control problems. This makes it difficult to align the concepts of RL with those of LLMs. On the other hand, most LLM literature lacks an in-depth discussion of RL techniques, often omitting related details. As a result, we find ourselves in an awkward middle ground between RL and LLMs, struggling to bridge the two fields.

The aim of this paper is to provide a comprehensive introduction to RL from the perspective of LLM training. We begin by introducing the basic concepts and algorithms of RL using terminology familiar to LLM researchers, and illustrate them through concrete LLM training examples. We then discuss recent advances in RL for LLMs, including improved reward modeling, advantage estimation, training efficiency, and policy optimization. Beyond standard LLM alignment, we further introduce how RL is applied to enhance reasoning capabilities, support long-horizon decision-making in LLM-based agents, and optimize multimodal understanding and generation models. Finally, we summarize promising future directions and provide an overview of commonly used systems and datasets.

References

- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>.
- Deepseek. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- OpenAI. Learning to reason with llms, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial intelligence*, 299:103535, 2021.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (2nd ed.)*. The MIT Press, 2018.
- Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence, 2025. URL <https://arxiv.org/abs/2507.20534>.