Reinforcement Learning: Concepts, Techniques, and Applications to Large Language Models

Sabína Ságová

Abstract

Reinforcement Learning (RL) is a foundational area of machine learning that empowers agents to make sequential decisions through trial-and-error interactions with an environment. Unlike supervised learning, which relies on labeled datasets, RL agents learn optimal behavior by maximizing cumulative reward signals. This learning paradigm is particularly powerful in domains involving uncertainty and long-term planning, such as robotics, control systems, and increasingly, the training and alignment of large language models (LLMs) like ChatGPT. This paper outlines core RL principles, summarizes key algorithmic approaches, and explores how RL—particularly Reinforcement Learning from Human Feedback (RLHF)—is reshaping the behavior of LLMs.

1. Introduction

Reinforcement Learning offers a framework for learning via interaction, where an agent learns to maximize a long-term reward signal through feedback from the environment. It has gained increasing relevance in generative AI, where aligning outputs with human expectations remains a critical challenge. RL provides a mechanism for continuous, feedback-driven refinement of model behavior.

2. Theoretical Foundations

At the core of RL lies the *Markov Decision Process (MDP)*, defined by:

- A set of states S
- A set of actions A
- A transition probability $p(s', r \mid s, a)$
- A reward function R
- A discount factor $\gamma \in [0, 1]$

At each timestep t, the agent observes state S_t, chooses action A_t, receives reward R_{t+1}, and transitions to state S_{t+1}, where future outcomes depend only on the current state and action—the Markov property.

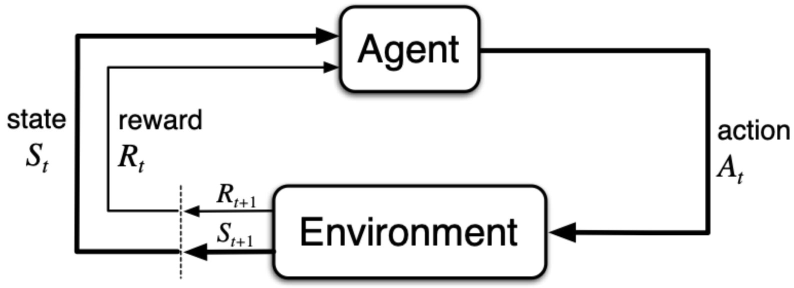

Figure 1: The agent–environment interaction in a Markov decision process.

3. Key Algorithms

Several families of RL algorithms are foundational:

- **Dynamic Programming**: Requires full environment knowledge (e.g., value iteration).
- **Monte Carlo Methods**: Learn from episodes, suitable for unknown environments.
- **Temporal Difference (TD) Learning**: Combines DP and Monte Carlo; includes SARSA and Q-Learning.
- **Policy Gradient Methods**: Optimize policies directly, especially useful in continuous spaces.
- **Deep Reinforcement Learning (DRL)**: Uses neural networks to approximate policies or value functions (e.g., DQN, PPO).

4. Reinforcement Learning in LLMs

Reinforcement Learning has become central to training aligned, safe large language models. Traditional supervised training cannot always capture the complexity of human preferences. This gap is addressed by **Reinforcement Learning from Human Feedback (RLHF)**.

The RLHF pipeline typically involves:

1. Generating multiple outputs from the model.
2. Ranking or scoring them using human evaluators.
3. Training a reward model to predict preference scores.
4. Fine-tuning the model via Proximal Policy Optimization (PPO) to maximize the predicted reward.

Example: A prompt such as “Explain black holes to a 12-year-old” may yield several completions. Human annotators rank them based on clarity and tone. The model then learns to produce outputs similar to the top-ranked completions.

5. Recent Developments

The Kimi k1.5 model (Kimi Team, 2025) exemplifies the practical benefits of reinforcement learning for language models. By combining supervised pretraining with reward-based fine-tuning, Kimi k1.5 improves response relevance and achieves state-of-the-art reasoning performance across multiple benchmarks. Its simplified, scalable RL approach—eschewing complex elements like value functions and Monte Carlo tree search—demonstrates how reinforcement learning can be effectively applied to real-world LLM use cases while maintaining robustness.

6. Conclusion

Reinforcement Learning, and particularly RLHF, provides a critical toolkit for aligning LLMs with human values and expectations. As generative models evolve in capability and scope, RL will remain essential in shaping their behavior, improving safety, and fostering human-centered AI.

References

1. Sutton, R. S., & Barto, A. G. (2020). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. `http://incompleteideas.net/book/RLbook2020.pdf`

2. Wang, S., et al. (2025). Reinforcement Learning Enhanced LLMs: A Survey, arXiv:2412.10400v3.

3. Kimi Team. (2025). Kimi k1.5: Scaling Reinforcement Learning with LLMs, arXiv:2501.12599v2.