# Reinforcement Learning from Human Feedback

A short introduction to RLHF and post-training focused on language models.

Nathan Lambert

16 April 2025

**Abstract**

Reinforcement learning from human feedback (RLHF) has become an important technical and storytelling tool to deploy the latest machine learning systems. In this book, we hope to give a gentle introduction to the core methods for people with some level of quantitative background. The book starts with the origins of RLHF – both in recent literature and in a convergence of disparate fields of science in economics, philosophy, and optimal control. We then set the stage with definitions, problem formulation, data collection, and other common math used in the literature. The core of the book details every optimization stage in using RLHF, from starting with instruction tuning to training a reward model and finally all of rejection sampling, reinforcement learning, and direct alignment algorithms. The book concludes with advanced topics – understudied research questions in synthetic data and evaluation – and open questions for the field.