Training Large Language Models and Teaching Them To Reason via Reinforcement Learning

compiled by D. Gueorguiev, Jan 25, 2025

# Reinforcement Learning Policy Optimization for Chat GPT

(1)

# References

[1] [Reinforcement Learning from Human Feedback, Nathan Lambert, Louise Castricato, Leandro von Werra, Alex Havrila, HugginFace article, Dec 9th, 2022](https://huggingface.co/blog/rlhf)

[2] [A Little Bit of Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training focused on language models. Nathan Lambert, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/A_Little_Bit_of_Reinforcement_Learning_from_Human_Feedback_Lambert_2025.pdf)

[3] [Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Rafel Rafailov et al, Stanford U., 2023](https://github.com/dimitarpg13/deep_learning_for_image_processing/blob/main/literature/articles/rlhf/Direct_Preference_Optimization-Your_Language_Model_is_Secretly_a_Reward_Model_Rafailov_Stanford_2023.pdf)

[4] [Robust Preference Optimization through Reward Model Distillation, Adam Fisch et al, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Robust_Preference_Optimization_through_Reward_Model_Distillation_Fisch_2024.pdf)

[5] [Training Language Models to Self-Correct via Reinforcement Learning, Aviral Kumar et al, Google DeepMind, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Training_Language_Models_to_Self-Correct_via_Reinforcement_Learning_Kumar_2024.pdf)

[6] [Vanishing Gradients in Reinforcement finetuning of Language Models, Noam Razin et al, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Vanishing_Gradients_in_Reinforcement_Finetuning_of_Language_Models_Razin_2024.pdf)

[7] [Scaling Reinforcement Learning with LLMs, Technical Report, Kimi, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Scaling_RL_with_LLM_Kimi_k1.5.pdf)

[8] [DeepSeek R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, DeepSeek AI, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeek_R1-Incentivizing_Reasoning_Capability_in_LLMs_via_RL.pdf)

[9] [DeepSeek Prover v1.5 Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search, H. Xin et al, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeek-Prover-V1.5-Harnessing_Proof_Assistant_Feedback_for_Reinforcement_Learning_and_Monte-Carlo_Tree_Search.pdf)

[10] [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, Z. Shao et al, DeepSeek AI, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/DeepSeekMath-Pushing_the_Limits_of_Mathematical_Reasoning_in_Open_Language_Models_Shao_2024.pdf)

[11] [Discussion on Reinforcement Learning techniques for training ChatGPT with Ajay Shenoy](https://www.linkedin.com/posts/bastyajayshenoy_chatgpt-was-trained-using-this-algorithm-activity-7287688159981682688-wMIX)

[12] [Linkedin discussion on DeepSeek and Kimi, Jim Fan](https://www.linkedin.com/posts/drjimfan_we-are-living-in-a-timeline-where-a-non-us-activity-7287125475280265217-dUT3?utm_source=share&utm_medium=member_desktop)  
[13] [Linkedin discussion on DeepSeek and Kimi, Jim Fan](https://www.linkedin.com/posts/drjimfan_that-a-second-paper-dropped-with-tons-of-activity-7287148978247290880-Gkd2?utm_source=share&utm_medium=member_desktop)  
[14] [Linkedin discussion on DeepSeek, Charles H Martin](https://www.linkedin.com/posts/charlesmartin14_revenge-of-the-quants-so-what-have-we-activity-7288268696253345792-zjQT)  
[15] [Linkedin discussion on DeepSeek, Charles H Martin](https://www.linkedin.com/posts/charlesmartin14_thanks-to-deepseek-no-one-is-sleeping-activity-7288988183332077569-9G6O)