Topic 3: SFT and RLHF

1. OpenAI blog overview of instruction-following and RLHF: https://openai.com/index/instruction-following/
2. Proximal Policy Optimization (PPO) paper: https://arxiv.org/abs/1707.06347
3. InstructGPT paper: https://arxiv.org/abs/2203.02155
4. Fine-Tuning Language Models from Human Preferences: https://arxiv.org/abs/1909.08593
5. OpenAI Spinning Up page on PPO (practical intro): https://spinningup.openai.com/en/latest/algorithms/ppo.html
6. Anthropic “Helpful and Harmless” RLHF writeup: https://www.anthropic.com/research/training-a-helpful-and-harmless-assistant-with-reinforcement-learning-from-human-feedback
7. Direct Preference Optimization (DPO) paper: https://arxiv.org/abs/2305.18290
8. How to run DPO in practice: https://huggingface.co/docs/trl/en/dpo_trainer
9. GRPO for reasoning LMs: https://arxiv.org/abs/2402.03300
10. TRL library overview (SFT, GRPO, DPO, reward modeling): https://huggingface.co/docs/trl/en/index
11. OpenRLHF framework (high-performance RLHF toolkit): https://github.com/OpenRLHF/OpenRLHF
12. RLHF Book (comprehensive guide to SFT, PPO, DPO, GRPO): https://rlhfbook.com/
13. OpenAI blog on RLHF for summarization: https://openai.com/index/learning-to-summarize-with-human-feedback/
14. RLVR paper: https://arxiv.org/abs/2506.14245
15. Prover-Verifier games improve legibility (verifiable reasoning signals): https://cdn.openai.com/prover-verifier-games-improve-legibility-of-llm-outputs/legibility.pdf
16. Hugging Face RLHF primer (illustrated explanation and links): https://huggingface.co/blog/rlhf
17. StackLLaMA tutorial (end-to-end RLHF recipe on LLaMA): https://huggingface.co/blog/stackllama