1. How do humans provide feedback to the LLM?
    a. The human feedback portion of Reinforcement Learning for Human Feedback (RLHF) is relatively new. For standardization, a *large* set of prompts are chosen, then the LMs are queried using these prompts. The responses are then analyzed by the humans. The rankings for the responses are and the LM used to provide the responses are then compared using an “Elo system”. I hadn’t heard of or had forgotten what and Elo system was, so this GitHub repo, this Medium article, and the Wikipedia article helped instruct me.
2. The article describes how to set up an LLM as a reinforcement learning problem. What are the states (or observation space), actions, policy, and reward?
    a. Policy is a language model digesting a prompt and returning a string of text.
    b. The action space is the collection of tokens mapped to the LM’s lexicon.
    c. The observation space is the collection of input token strings. Where L = the lexicon of the model, and s is the number of input token sequences, the size of the observation space is |L|^s.
    d. The reward is a “combination of the preference model and a constraint on policy shift” (Illustrating Reinforcement Learning from Human Feedback (RLHF)).
3. Why are there so few RLHF datasets available?
    a. The primary reason limiting the availability of datasets is collecting human feedback. “Generating well-written human text answering specific prompts is very costly, as it often requires hiring part-time staff” (Illustrating Reinforcement Learning from Human Feedback (RLHF)). Additionally, the cost of training the reward model isn’t cheap but would likely be too large a sum for academic research centers. Secondly, the human annotators can disagree. This would create some interesting results in the reward model when tuning it.