Andrew Taylor EN.705.743 12/01/25

Reading Assignment 13: SFT and RLHF

Q1.

Supervised Fine-Tuning is included in the RLHF pipeline because reinforcement learning alone is too unstable when applied directly to a raw pretrained model. Before RL can be effective, the model needs to operate within a reasonable region of the action space, producing coherent and task-relevant outputs that a reward model can meaningfully evaluate. SFT provides this foundation by aligning the model with high-quality demonstrations, giving RL a structured starting point rather than an unbounded search problem. It also reduces the likelihood of early reward hacking, since the model begins with behaviors that approximate human preferences. As a result, RL builds on a stable base rather than trying to shape behavior from scratch.

Q2.

A reward model can introduce several pitfalls, including misgeneralization from limited preference data, sensitivity to annotator bias, and a tendency to overfit superficial cues instead of the deeper qualities humans intend to reward. When the policy model optimizes against these imperfections, it may exploit blind spots in the reward function, producing output that appears highly rated but diverges from real human intent. This creates downstream alignment risks such as overconfidence, sycophancy, or strategically misleading behavior. Another challenge arises from distributional shift: as the policy improves, it begins generating outputs outside the distribution the reward model was trained on, reducing the reliability of its judgments. These issues can cause the RL stage to reinforce errors rather than correct them, degrading alignment over time.