

- 强化学习
 - RLHF
 - 奖励模型

强化学习

- 有监督微调通常采用交叉熵损失做为损失函数，目标是调整参数使得模型输出与标准答案完全相同，不能从整体上对模型输出质量进行判断。因此，模型不能适用自然语言多样性，也不能解决微小变化的敏感性问题。
- 强化学习则将模型输出文本作为一个整体进行考虑，其优化目标是使得模型生成高质量回复
- 根据智能体所学习的组件的不同，可以把智能体归类为：基于价值的智能体、基于策略的智能体和演员-评论员智能体。
 - 基于价值的智能体（**Value-based Agent**）显式地学习价值函数，隐式地学习策略。其策略是从所学到的价值函数推算得到的。
 - 基于策略的智能体（**Policy-based Agent**）则是直接学习策略函数。策略函数的输入为一个状态，输出为对应动作的概率。基于策略的智能体并不学习价值函数，价值函数隐式的表达在策略函数中。
 - 演员-评论员智能体（**Actor-critic Agent**）则是把基于价值的智能体和基于策略的智能体结合起来，既学习策略函数又学习价值函数都，通过两者的交互得到最佳的动作。
- 强化学习
 - 比有监督学习更可以考虑整体影响
 - 更容易解决幻觉问题
 - 可以更好的解决多轮对话的奖励累积问题

RLHF

- 两阶段：奖励模型训练和近端策略优化(PPO)

