1. 在模仿学习中，训练数据（专家经验）通常是有限的，仅有正向样本，而Agent在实际部署中可能由于没有充足的经验而面临专家经验中没有见过的观测，它将不知该如何处理。同时，使用监督学习使得Agent只能辅助专家行为，不知道真实的奖励函数，因此很难学习到好的效果。

**问题1：模仿学习中由于训练数据（专家经验）有限且缺乏负向样本，Agent在面对未见过的观测时难以处理，且由于无法直接获取真实奖励函数，难以在监督学习框架下实现最优策略。**

1. 对于序列观测问题，无法满足训练集和测试机分布一致的假设，因为后续状态和当前的动作分布有关，而Π（Agent）和Π※（专家经验）通常不一致，导致训练和测试集中的分布不同，同时，由于序列决策问题误差会积累，导致Agent难以学习到有效的完整策略。

**问题2：在使用监督学习解决序列决策问题中，由于Agent策略与专家策略不一致，导致训练集和测试集分布不同，加之误差在序列决策过程中积累，进一步增加了学习有效完整策略的难度。**

1. Agent难以从次优的专家经验中学习到最优解，例如：使用GPT-4o当专家经验，LLaMA无法从中学习到优于GPT4o的策略，GPT4o也没有办法从自己的经验中获得提升。

**问题3：在模仿学习中，由于次优专家经验的局限性，Agent难以学习到超越专家的策略，例如，LLaMA无法从GPT-4o的经验中获得优于其策略的提升，而GPT-4o也无法通过自身经验实现进一步优化。**

针对问题1和3，本文通过让Agent从专家经验学习到初始策略后，和环境进行交互的办法获得Π策略在实际中的真实观测分布，训练：安全性和有用性奖励模型对Agent采样轨迹进行打分，再训练新的策略Π’，同时在交互过程中加入探索机制，使得Agent学习到优于专家经验的策略。