从经典PPO到PPO-RLHF（一）

东川路第一可爱猫猫虫

主要内容

- Deep Reinforcement Learning from Human Preferences

  如果强化学习没有了环境的reward
- reward hacking
- reward model
- 建立从经典RL到LLM的映射

感谢

阿萨李逍遥张小凡

梦泽5867

修改呢称消耗枚硬币

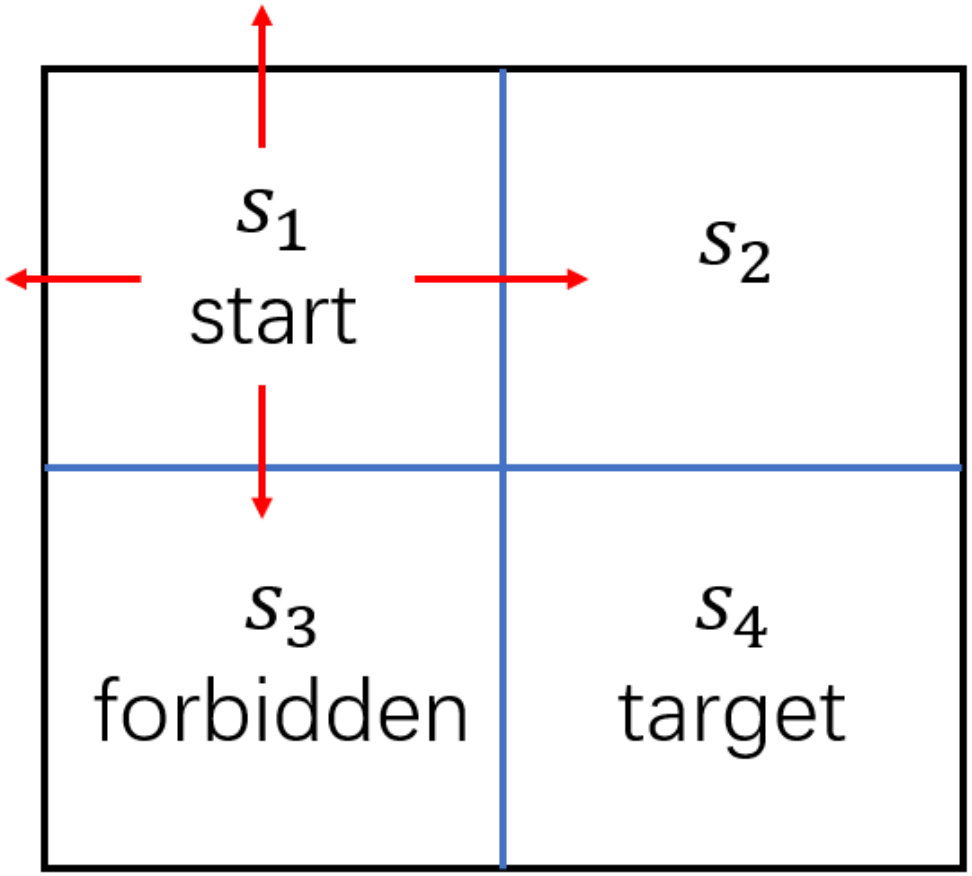
迷途小白哟

W0ND3RFULHE4VEN

RLHF的先驱

Deep Reinforcement Learning from Human Preferences

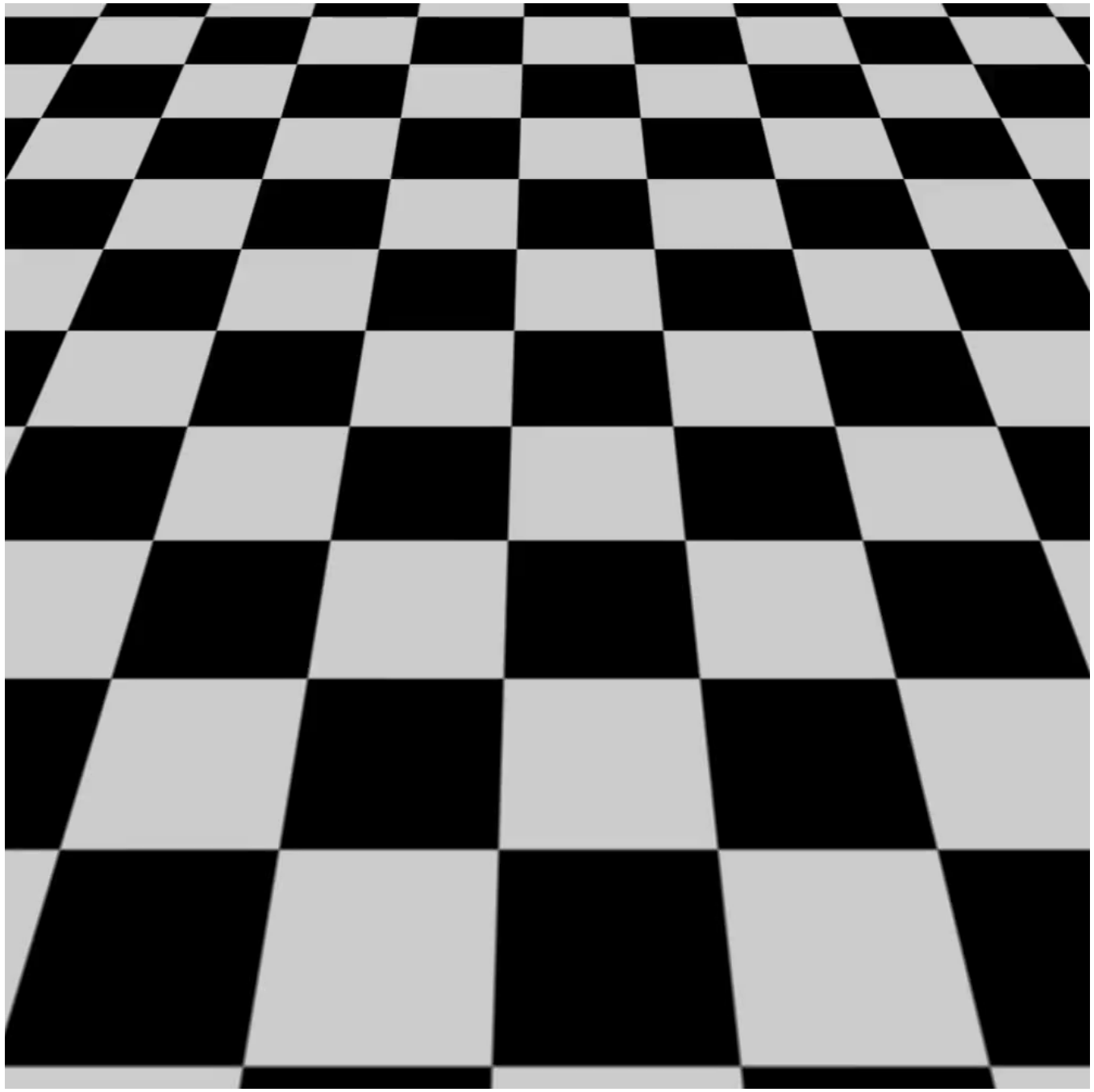
- 通过人类的偏好来学习奖励函数
  用这个奖励函数训练强化学习智能体
- 使用 Bradley-Terry 模型量化人类偏好
- 用人类偏好替代或补充环境奖励



- 当智能体采取动作，完成状态转移
  环境会给智能体奖励
- 但如果任务本身没有分数呢
  没有了环境自带的奖励

没有环境自带的简单奖励

- 有些任务的奖励
  - 可能很难定义或精确描述
- 比如训练一个小机器人去擦桌子或者煎鸡蛋
  - 它完成的好坏很难用一个简单的奖励函数来评判
- 可不可以设计一个简单的目标函数来近似
  - 我们想要智能体学会的行为
  - 代理目标函数
- 比如我们想要训练小机器人学会后空翻
  - 尝试定义奖励=高度变化速度+关节转动速度



reward hacking

- 智能体的偷懒

  小机器人通过捷径拿到了高分

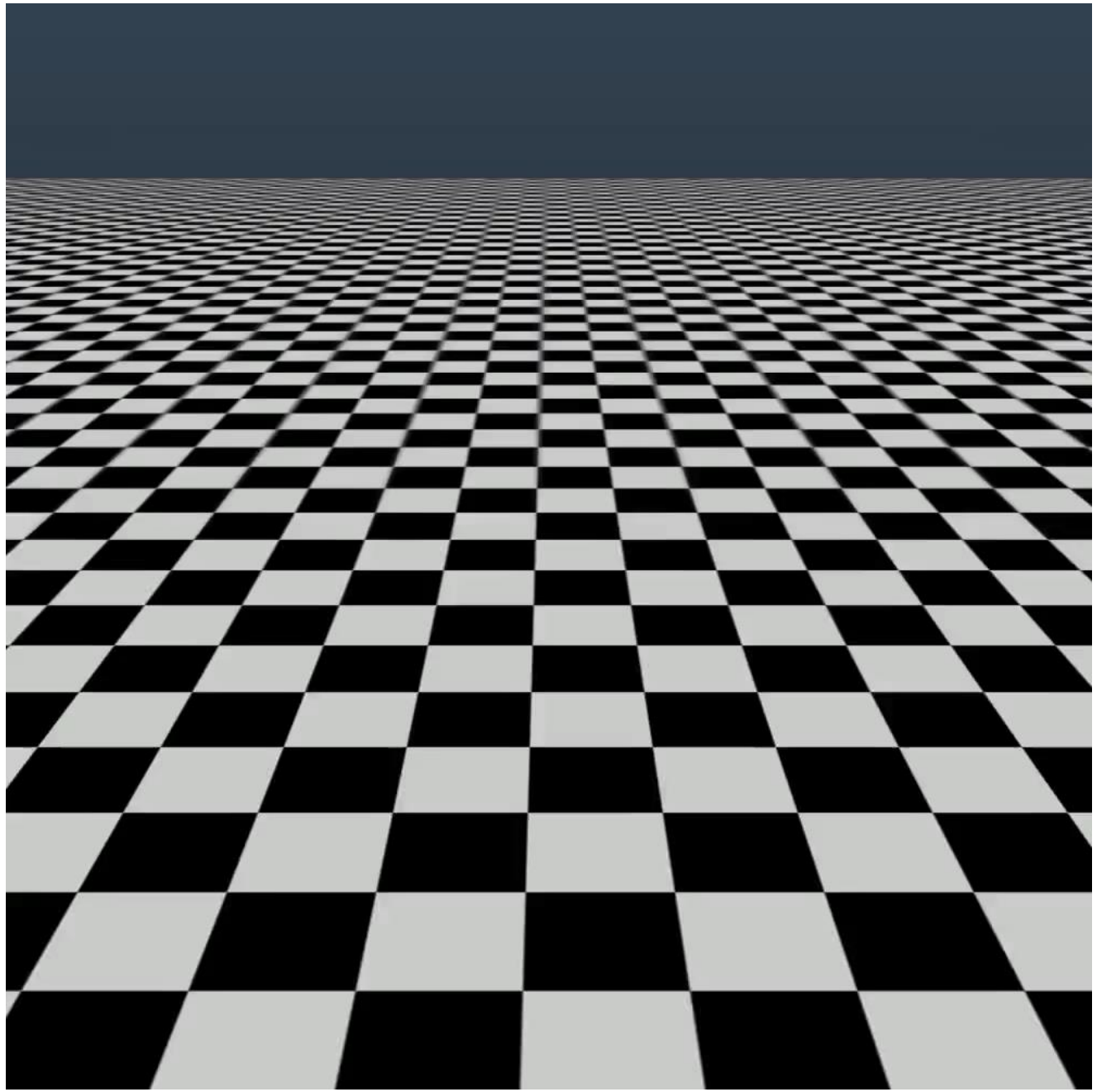
  但它做的不是后空翻而是蹭地
- reward hacking奖励作弊　misaligned目标错位
- Anthropic在训练ClaudeSonnet3.7的时候

  我们希望训练智能体学会写代码

  但智能体会通过作弊之类的方式来通过测试

reward model

- reward hacking的本质
  - 很多对人类来说可以一眼看懂的任务
  - 难以用简单的公式和代码来表示
- 让人类来评判任务的完成情况
  - 给各个动作分别打分
  - 给定两个动作，选择优劣
- 收集人类的二选一数据
  - 训练出一个神经网络来为智能体提供奖励
  - 这个神经网络我们称为reward model



把RL应用到LLM里

- 需要建立从**经典RL概念**到**大模型实体**的映射

  state状态

  state space状态空间

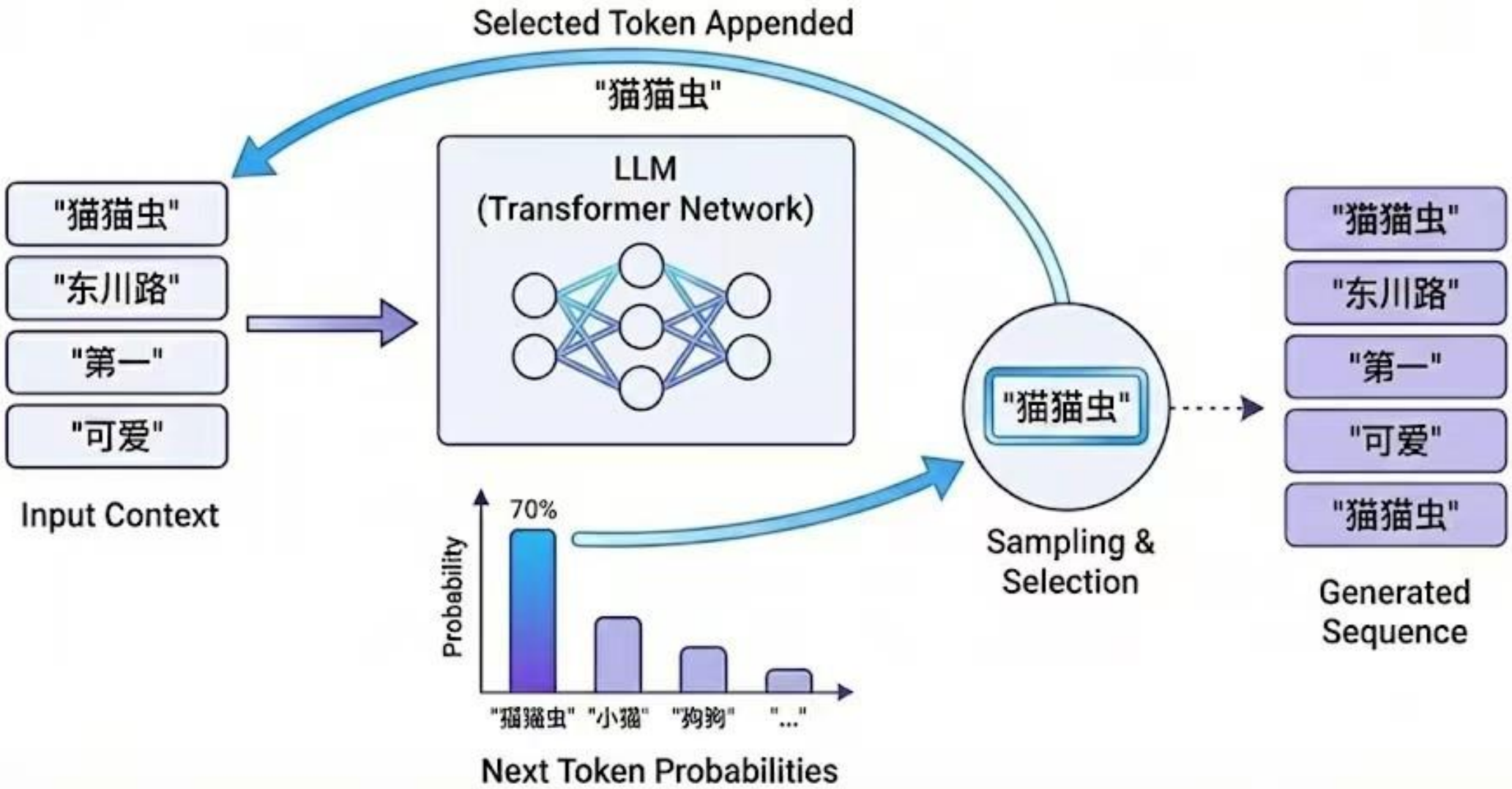
  action行动

  action space of a state状态的行动空间

  transition dynamics状态转移

  reward function奖励函数

Autoregressive Token Generation Process in LLMs



策略 动作 状态

- LLM里的policy

  LLM

  它决定了给定上下文x的情况下生成序列y的概率

- LLM里的action

  a_t是t时刻大模型选择生成的token

  也可以是序列或者片段

- LLM里的state

  t时刻的states_t

  初始提示词x与迄今为止生成的所有token

  $s_t = (x, a_{1:t-1})$

状态转移 奖励

- LLM里的transition dynamics

  确定性的

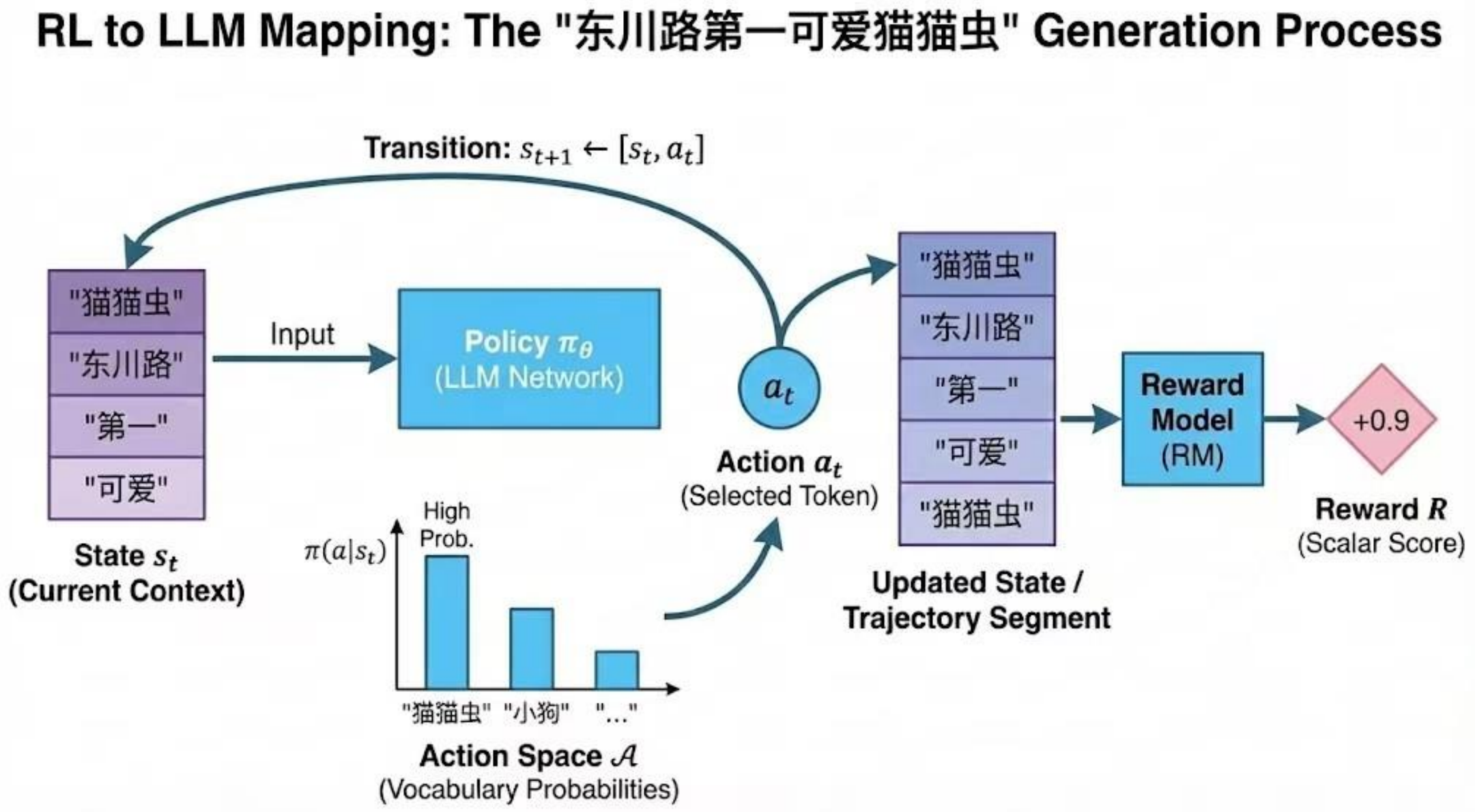
  新状态s_{t+1}一定是旧状态s_t与行动a_t的拼接

- LLM里的reward

  奖励可以是token级的 sequence级的 segment级的

  RLHF里奖励

  生成序列之后，reward model给出一个标量奖励



我们可以把人类反馈应用到LLM里

Training language models to follow instructions with human feedback

- 下一个视频我们将以InstructGPT为例
  讲解RLHF的完整流程
  如何用人类反馈来让LLM与人类偏好对齐