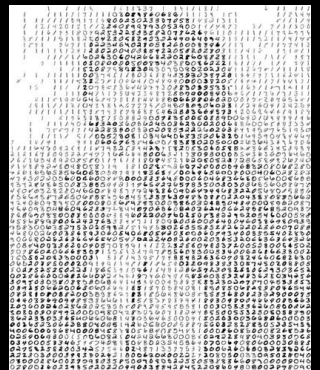
Toward Efficient Exploration by Large Language Model Agents

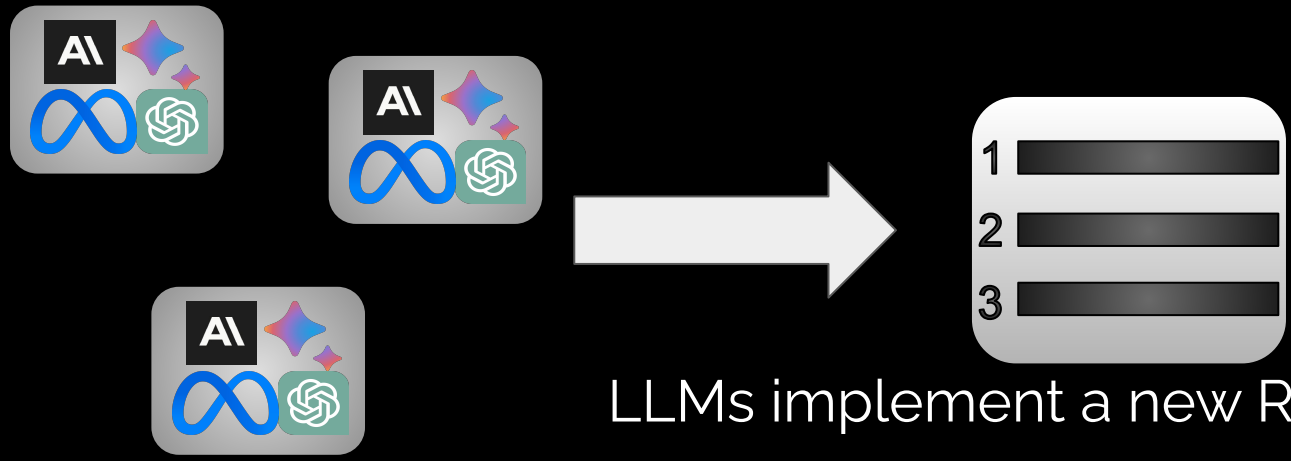
Dilip Arumugam & Tom Griffiths

dilip.a@cs.princeton.edu



AI LAB
Princeton Laboratory
for **Artificial Intelligence**

An Alternative Agent Design Principle



LLMs implement a new RL algorithm

An old RL algorithm implemented by LLMs

Bayesian RL ⇔ environment is a random variable

$$\mathbb{P}(\mathcal{M} \in \cdot \mid H_t)$$

Epistemic uncertainty in environment drives uncertainty in optimal behavior

[Bellman & Kalaba; 1959], [Duff; 2002], [Ghavamzadeh *et al.*; 2015]

Posterior Sampling for Reinforcement Learning (PSRL)

Intuition

Learn optimal policy by identifying $\mathcal{M}$

Implementation

Thompson Sampling over MDP models

At each episode k in [K]:

- Sample MDP $M_k \sim \mathbb{P}(\mathcal{M} \in \cdot \mid H_k)$
- Execute optimal policy in environment $\pi^{(k)} = \pi^{\star}_{M_k}$
- Update history and induce posterior $\mathbb{P}(\mathcal{M} \in \cdot \mid H_{k+1})$
  $H_{k+1} = H_k \cup \tau_k$

Provably-efficient Bayesian RL in tabular MDPs (and more)

[Strens; 2000], [Osband, Van Roy, & Russo; 2013], [Osband & Van Roy; 2014], [Abbasi-Yadkori & Szepesvári; 2014], [Osband & Van Roy; 2017], [Agrawal & Jia; 2017], [Ouyang *et al.*; 2017], [Lu & Van Roy; 2019], [**DA** & Van Roy; NeurIPS 2022]

A LLM-Based Implementation of PSRL

At each episode k in [K]:

- Sample MDP $M_k \sim \mathbb{P}(\mathcal{M} \in \cdot \mid H_k)$
- Execute optimal policy in environment $\pi^{(k)} = \pi^{\star}_{M_k}$
- Update history and induce posterior $\mathbb{P}(\mathcal{M} \in \cdot \mid H_{k+1})$

$$H_{k+1} = H_k \cup \tau_k$$

A LLM-Based Implementation of PSRL

"Posterior" Sampling LLM

Given current knowledge and uncertainty, draw one hypothesis

Optimal Policy LLM

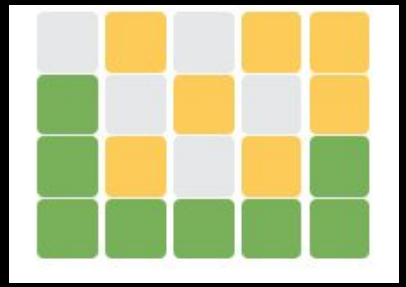
Given hypothesis and current state, take optimal actions

Approximate "Posterior" LLM

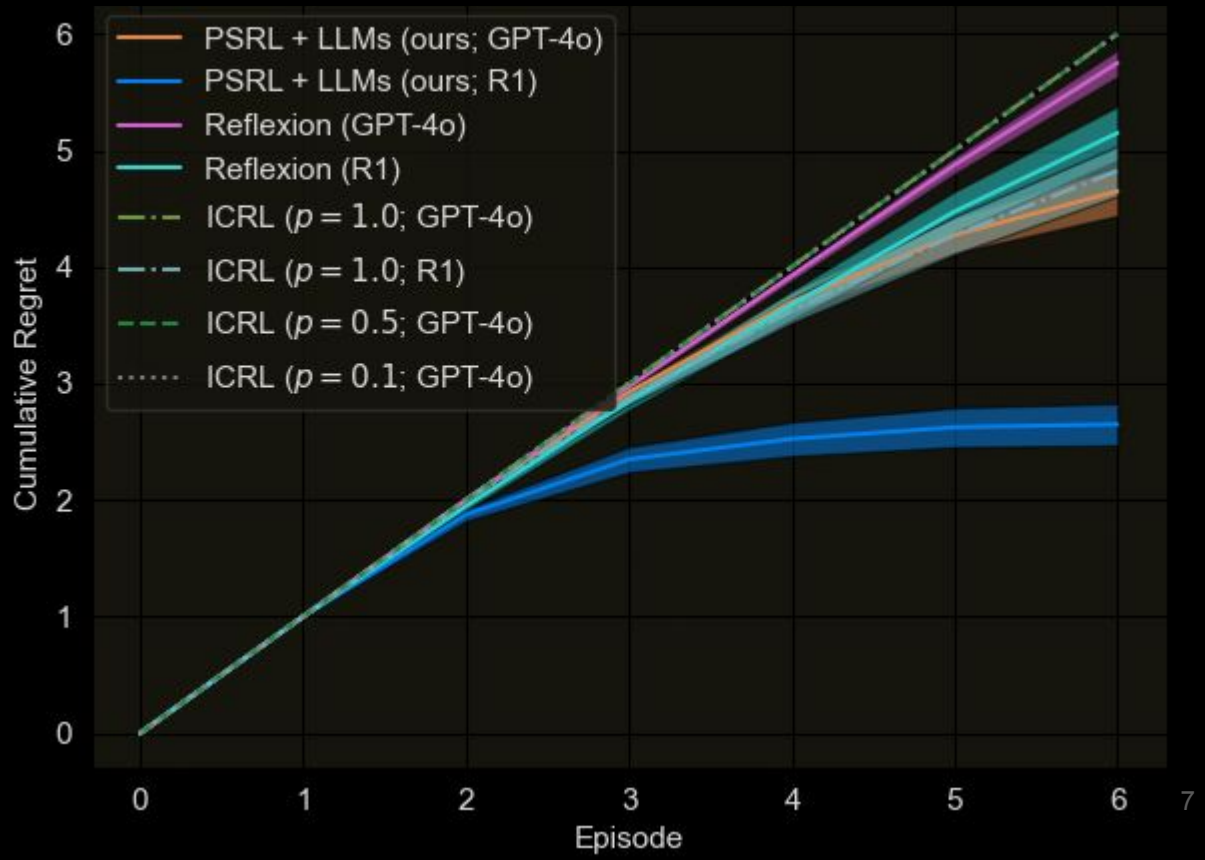
Given current beliefs and trajectory, update knowledge & uncertainty

Wordle

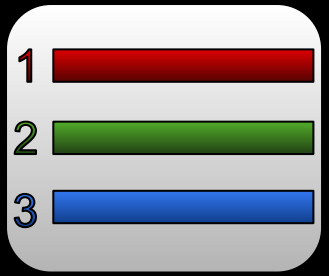
5-Letter Dictionary Word



- Letters A-Z
- Binary success reward
- Word chosen uniformly
- No repeating letters
- K = 6 episodes
- Wordle-style feedback



An old RL algorithm implemented by LLMs



Choosing PSRL retains LLM capabilities & inherits efficient exploration

As LLMs become more capable, viable task space grows

Promising initial results for information-directed exploration

[Friday Afternoon Poster Session]