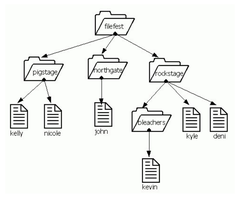
Reflect-RL: Two-Player Online RL Fine-Tuning for LMs

Runlong Zhou[1] Simon S. Du[1] Beibin Li[2]

[1]University of Washington [2]Microsoft Research, Redmond

Motivation



Given a **repository**
- Code
- Document
- Database files
- ……

- What is the fallback method for X if there is no Y?
  - **Document retrieval**
- Draw a map colored by employee salary.
  - **Coding, database query**

Challenges

Privacy
- Proprietary data
- Local model

Cost
- Smaller model
- Shorter prompt
- Faster response

Abstraction

Autonomous exploration in a file system given natural language queries
- **Locate**, modify and execute the correct files

We focus on file locating in this work

Fine-tune a local, small language model
- Privacy, cost

Contextual **reinforcement learning**
- Find **shortest path** given context (query)

Markov Decision Processes

- State $s = (s_1, s_2, \dots, s_L) \in \mathcal{S}$ ← Tokens
- Action $a = (a_1, a_2, \dots, a_K) \in \mathcal{A}(s)$
- Deterministic reward $r(s, a) \in [-1, 1]$
- Deterministic transition $\mathcal{T}(s, a) \in \mathcal{S}$
- Context (task) distribution μ
- SFT dataset $\mathcal{D}$ for **instruction following**

Reflection Agent

- Get reflection $R(s)$ from GPT-4 into $\mathcal{D}$
- SFT a local reflect agent $\hat{R}_\phi$

$$\mathcal{L}_{\mathrm{R}}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} -\log \hat{R}_\phi(R_{i,j}|s_i, R_{i,1:j-1})$$

- Freeze during RLFT

Action Generation

- Enumerate function
  $$\alpha\big(\mathcal{A}(s)\big) = (1, a_1; 2, a_2; \dots)$$
- Output only choice number
  - Reduce lm_head size
- Time complexity: $\Theta\big(|s|^2 + \sum_{a\in\mathcal{A}(s)}|a|^2\big)$
  - $|\mathcal{A}(s)|$ times faster than SayCan!

Policy Agent

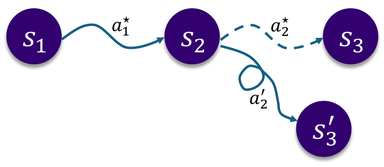
- SFT loss

$$\mathcal{L}_{\mathrm{P}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_i} -\log \pi_\theta(a_{i,j}|s_i, R_i, \alpha_i, a_{i,1:j-1})$$

- RLFT using REINFORCE

Negative Data

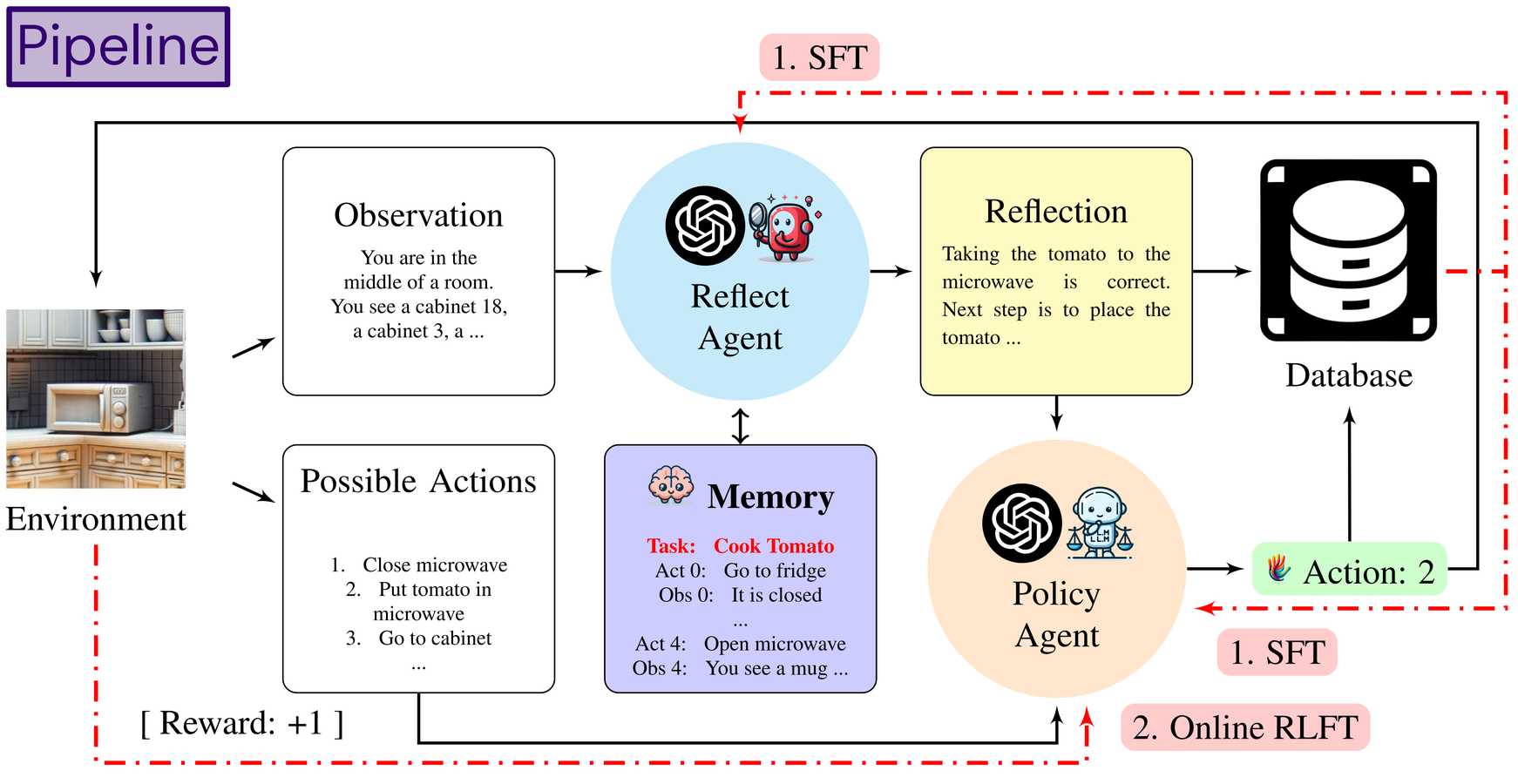
- Reflection agent **unable to correct errors** using only optimal data in $\mathcal{D}$
- **Perturb** each action on an optimal trajectory and tell GPT-4 it is sub-optimal to get negative reflection data



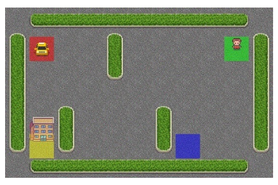
Curriculum Learning

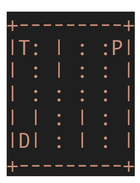
- Gradually increase depth for AutoExplore
- Extra pickup reward for Taxi

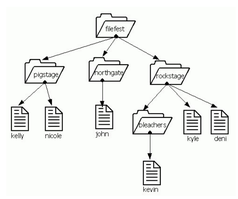
Pipeline



Benchmarks







AutoExplore

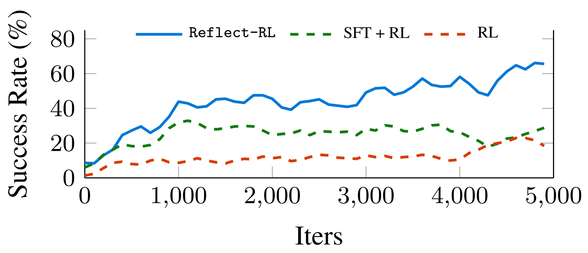
OpenAI Gym's **Taxi** in text
- More challenges
  - Invalid pickup / dropoff kills the game
  - Bumping into wall kills the game

ALFWorld
- Navigate in a **text** environment to complete a given goal

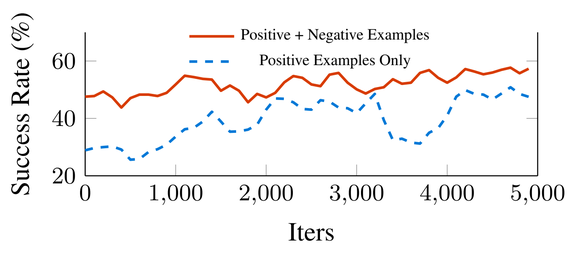
Results

1.56B model with Reflect-RL V.S. **7B** open-source models

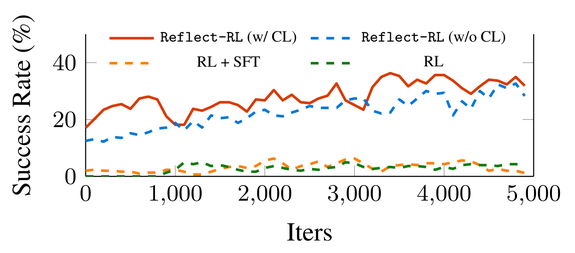
	Model	AutoExplore Depth 1	AutoExplore Depth 2	DangerousTaxi Pickup	DangerousTaxi +Dropoff*	ALFWorld
Open Source	Mistral 7B	34%	3%	7%	0%	0%
	Llama2 7B-chat	2%	1%	3%	0%	0%
	Orca-2 7B	6%	1%	1%	0%	0%
SFT Only	GPT-2 XL 1.56B	4%	9%	7%	0%	0%
RLFT Only	GPT-2 XL 1.56B	12%	3%	2%	0%	0%
SFT+RLFT (w/o reflection)	GPT-2 XL 1.56B	20%	4%	6%	0%	66%
SFT+RLFT (w/o negative)	GPT-2 XL 1.56B	33%	12%	-	-	-
Reflect-RL (Ours)	GPT-2 XL 1.56B	**36%**	**17%**	**58%**	**29%**	**74%**



Ablation on reflection: Success rate of Taxi pickup



Ablation on negative data: Success rate of AutoExplore



Ablation on curriculum learning (and reflection): Success rate of Taxi dropoff

vectorzh@cs.washington.edu