upstage / June 26th 2026

How Do LLMs Know They Are Wrong?

Exploration and the Dual Role of Self-Distillation

Jeonghye Kim

PhD student @ KAIST

Research Intern @ Microsoft Research

Short Bio

Education

- Ph.D. candidate in Electrical Engineering, KAIST, 2024.03-Current.
- M.S. in Electrical Engineering, KAIST, 2022.03-2024.02.
- B.S in Computer Science, KAIST, 2015.03-2022.02
- B.A. in Psychology, Bachelor's Degree Examination for Self-Education, 2018

Work Experience

Microsoft Research
2026.06-2026.08, Montreal

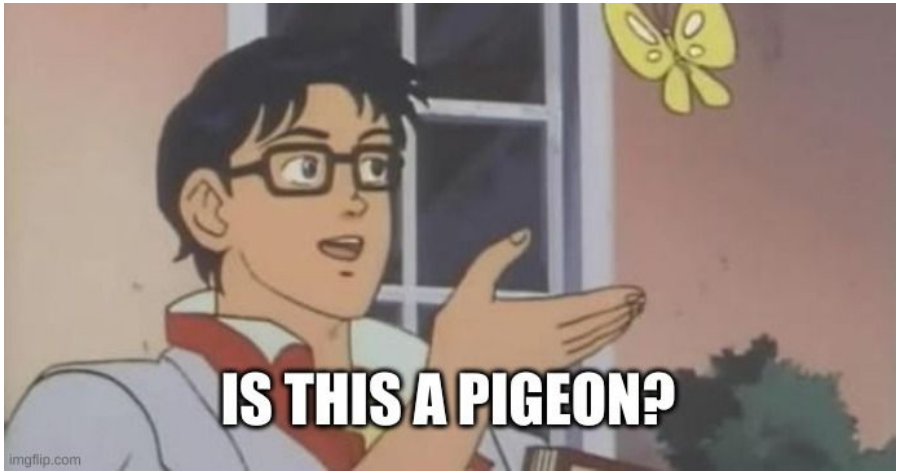
Microsoft Research
2025.09-2026.05, Shanghai

LG AI Research
2024.03-2024.11, Seoul

The Question: How does a model know it’s wrong?

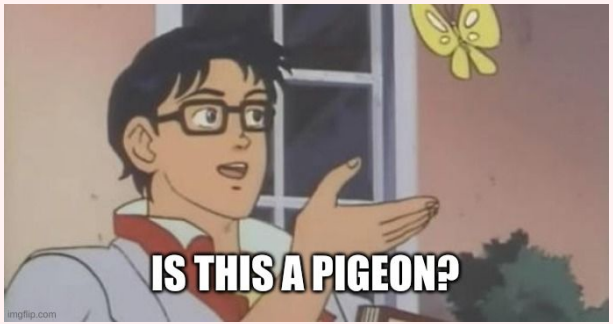
How Does a Model Know It’s Wrong?

We often ask whether a model can produce the right answer.
But today I want to ask a slightly different question:
when the model is wrong, how does it know?

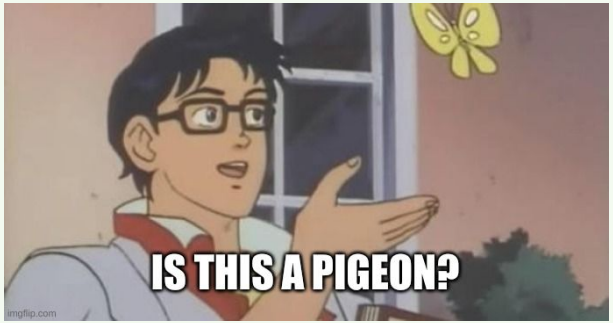


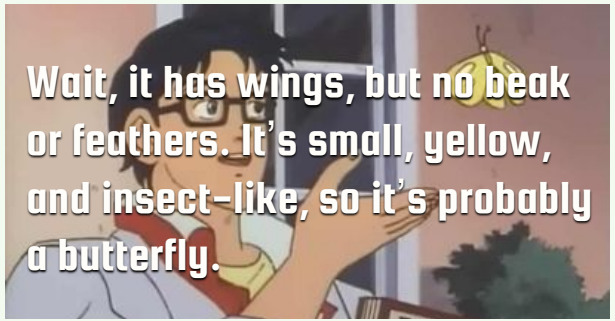
Why this question matters

Reasoning is not just answer generation.
Reasoning is also **error detection, correction, and exploration**.



Answer: PIGEON ❌





Answer: BUTTERFLY

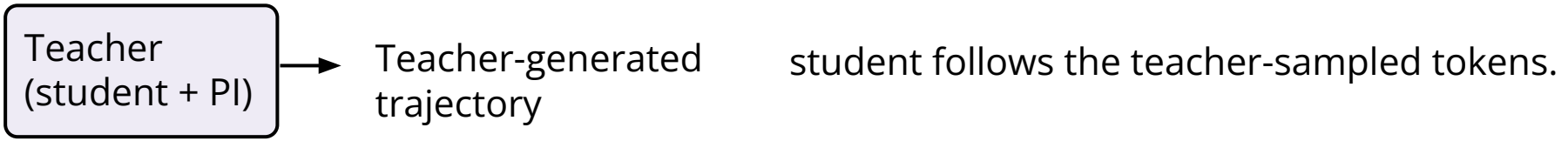
✅

What Is Self-Distillation?

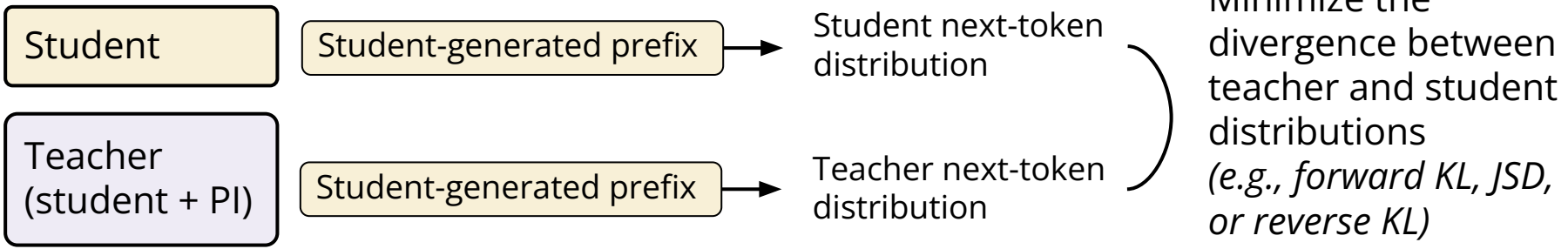
*We focus on **self-distillation conditioned on privileged information (PI)**.

What Is Self-Distillation?

Off-Policy Self-Distillation (SFT)



On-Policy Self-Distillation



When the World Provides the Error Signal

World-interactive reasoning

TextWorld

Welcome!

You are in the middle of the room. Looking around you, you see a diningtable, a stove, a microwave, and a cabinet.

Your task is to:
Put a pan on the diningtable.

> goto the cabinet

You arrive at the cabinet. The cabinet is closed.

> open the cabinet

The cabinet is empty.

> goto the stove

You arrive at the stove. Near the stove, you see a pan, a pot, a bread loaf, a lettuce, and a winebottle.

> take the pan from the stove

You take the pan from the stove.

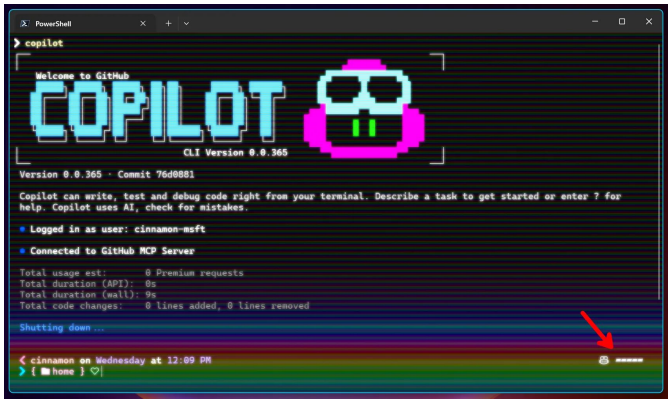
> goto the diningtable

You arrive at the diningtable.

> put the pan on the diningtable

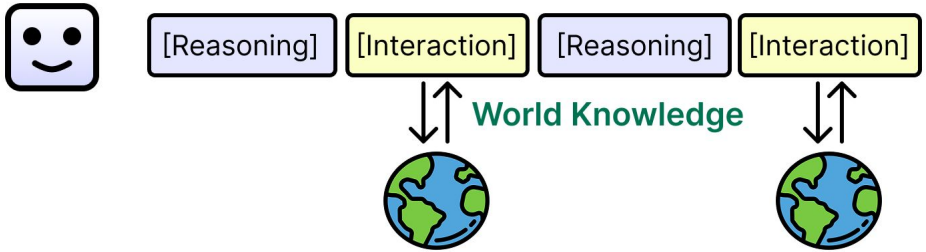
You put the pan on the diningtable.

Embodied



https://x.com/voooooogel/status/1847631721346609610
https://developer.microsoft.com/blog/making-windows-terminal-awesome-with-github-copilot-cli

World-interactive reasoning



- Facing an **unfamiliar environment,** the agent must **try many new and diverse things**.
- It then improves incrementally from **world feedback**, learning how to solve the task in this new environment.

- Self-distillation drives both **exploration** and **world knowledge internalization**.

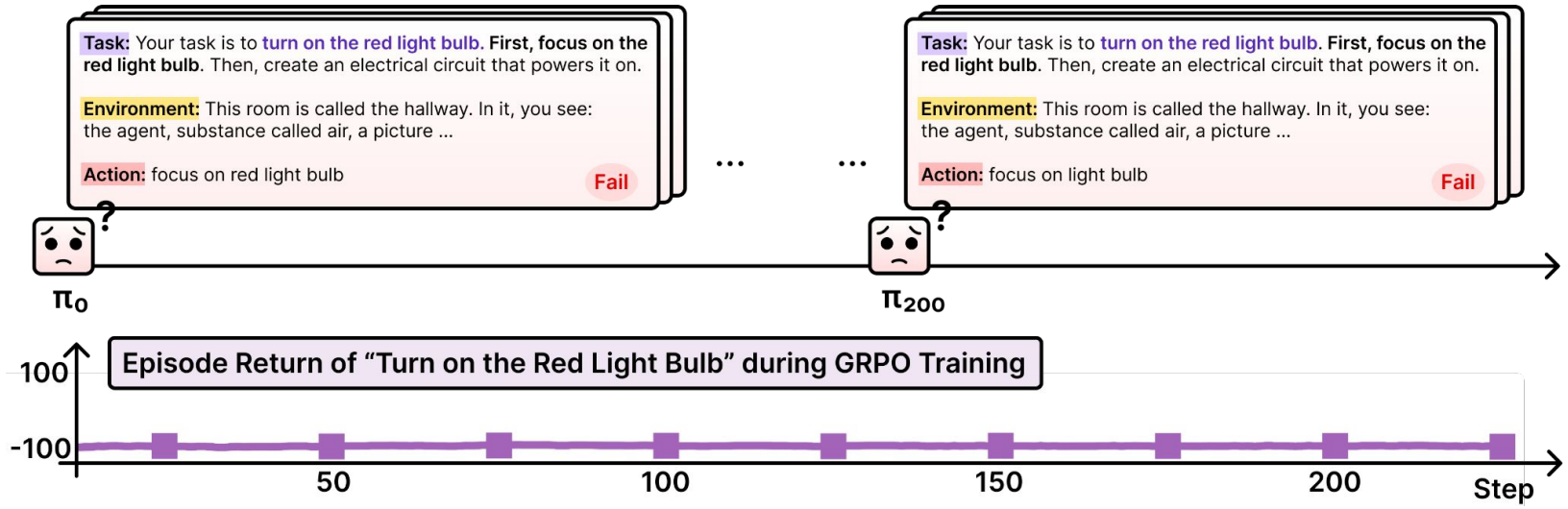
EMPO^2: Exploratory Memory-Augmented LLM Agent via Hybrid On- and Off-Policy Optimization

Zeyuan Liu* , **Jeonghye Kim*** , Xufang Luo , Dongsheng Li , Yuqing Yang

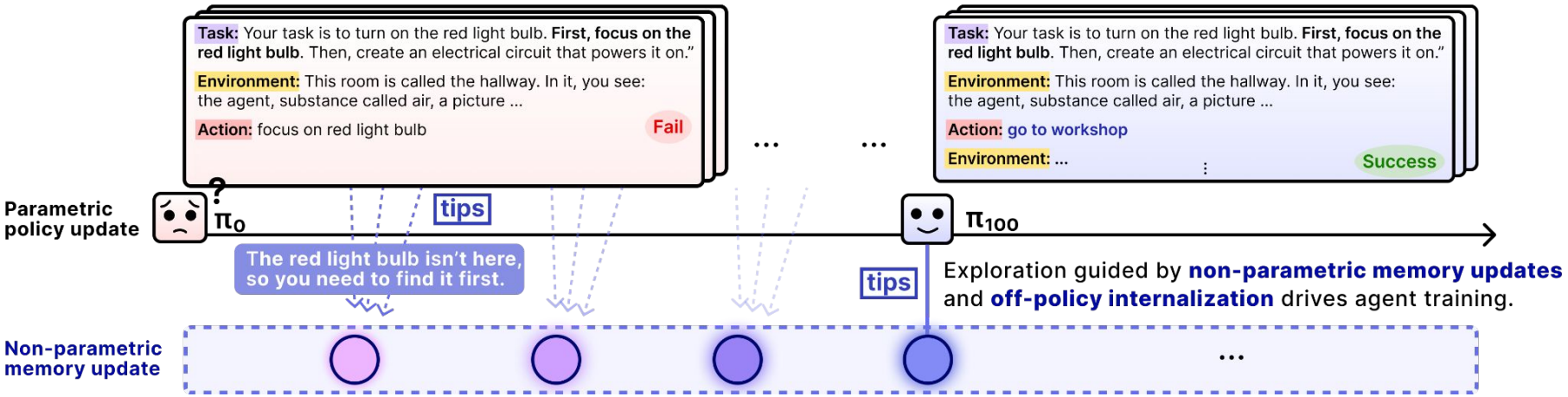
ICLR'26

The exploration problem in long-horizon agents

Long horizon, sparse rewards, bad actions compound, exploration is expensive
In GRPO training, **experience from previous attempts is not sufficiently reused in later rollouts.**



EMPO^2: Turning Past Experience into Future Tips



After each trajectory, the model writes memory tips.
In future rollouts, it retrieves relevant tips and conditions on them.

EMPO²: Using Tips to Correct and Explore

ScienceWorld <chemistry-mix-paint-secondary-color> task

[Without Memory]

Task: Your task is to use chemistry to create green paint. When you are done, focus on the green paint.
Observation: This room is called the hallway. In it, you see:
- the agent
- a picture
You also see:
- A door to the art studio (that is open)
...
Action: pour cup containing yellow paint in hallway in bowl

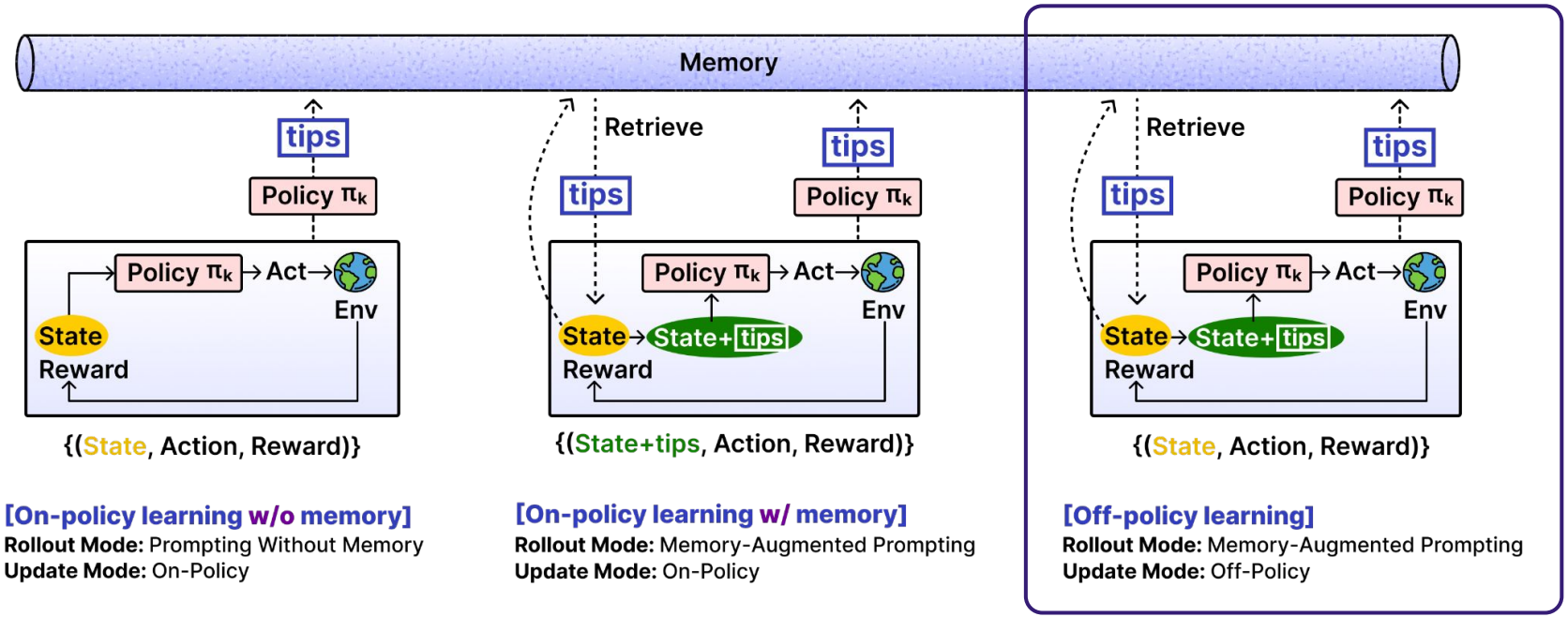
 # Task Failed

[With Memory-Augmented Prompting]

Task: Your task is to use chemistry to create green paint. When you are done, focus on the green paint.
Observation: This room is called the hallway. In it, you see:
- the agent
- a picture
You also see:
- A door to the art studio (that is open)
...
tips: Here are some memories you collected in your previous exploration:
Need to find a green pigment or mixture to create green paint.; At that timestep, the specific action your took was pour cup containing yellow paint in hallway in bowl; Eventually you got the score -100.0/100.
Need to find a green pigment or mixture to create green paint.; At that timestep, the specific action your took was open door to art studio; Eventually you got the score -100.0/100.
Action: go to art studio
Observation: You move to the art studio.
...

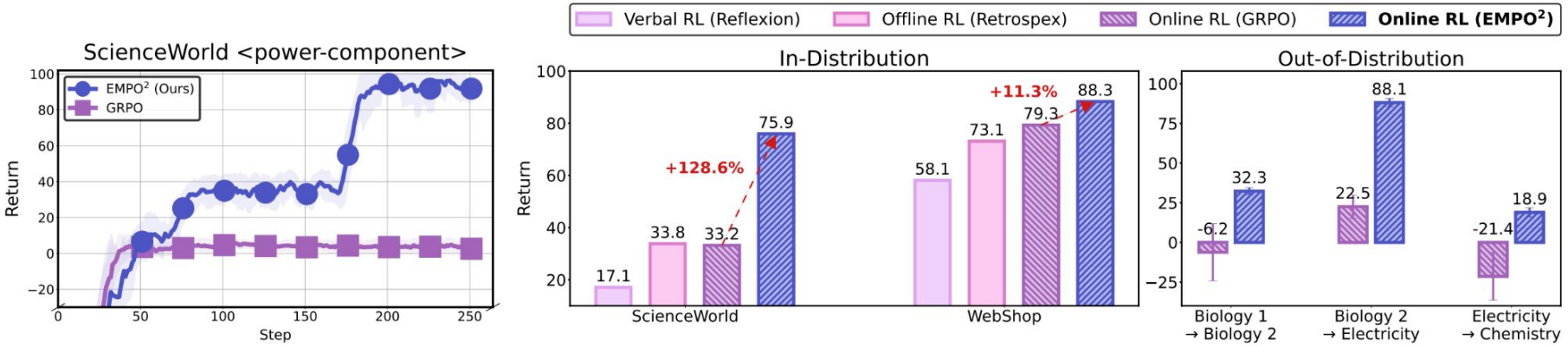
- The tips provide a summary of the previous rollout along with guidance regarding the actions taken.
- This guidance helps the **agent avoid repeating past failures** and **enables new action exploration.**

EMPO²: Distilling Tips into the Model



- We use off-policy learning to internalize tip-induced exploration gains into model parameters.
- The model rolls out with memory tips, then learns to reproduce useful behaviors without seeing those tips.

EMPO²: Results



- While GRPO converges to suboptimal performance, EMPO2 continues to improve and accomplish the task.
- In in-domain experiments, it adapts well to familiar environments, achieving 128.6% on ScienceWorld and 11.3% on Webshop improvements over GRPO. In OOD experiments, it also shows strong performance with few trials and no weight updates, indicating effective use of memory to explore unfamiliar environments.

On-Policy Self-Distillation in Practice: Cursor

May 18, 2026

Introducing Composer 2.5

7 min read

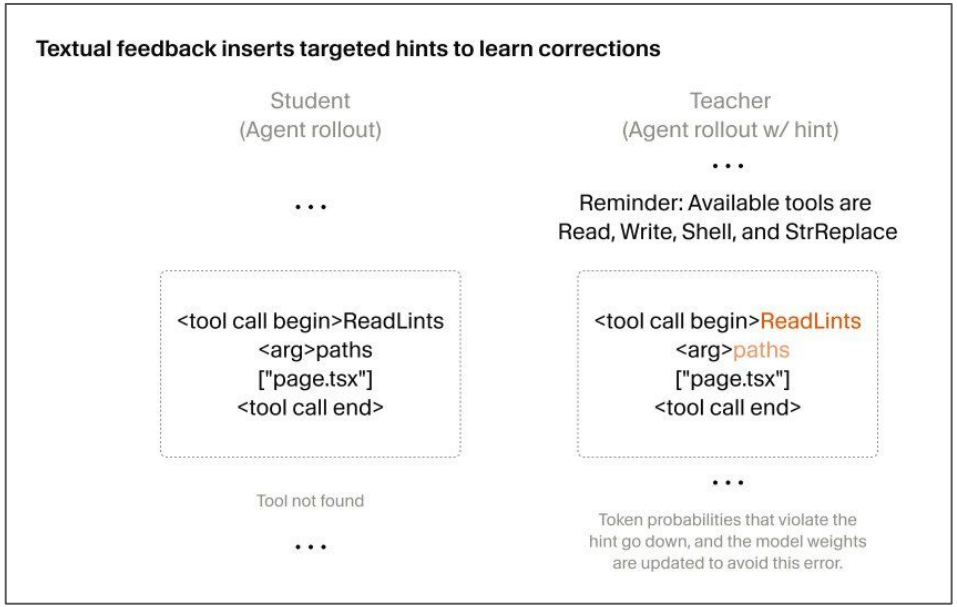
Table of Contents

- Training Composer 2.5
  - Targeted RL with textual feedback
  - Synthetic data
  - Sharded Muon and dual mesh HSDP
- Try Composer 2.5

Composer 2.5 is now available in Cursor.

It's a substantial improvement in intelligence and behavior over Composer 2. It is better at sustained work on long-running tasks, follows complex instructions more reliably, and is more pleasant to collaborate with.

	Composer 2.5	Opus 4.7	GPT-5.5	Composer 2
Terminal-Bench 2.0	69.3%	69.4%	82.7%	61.7%
SWE-Bench Multilingual	79.8%	80.5%	77.8%	73.7%



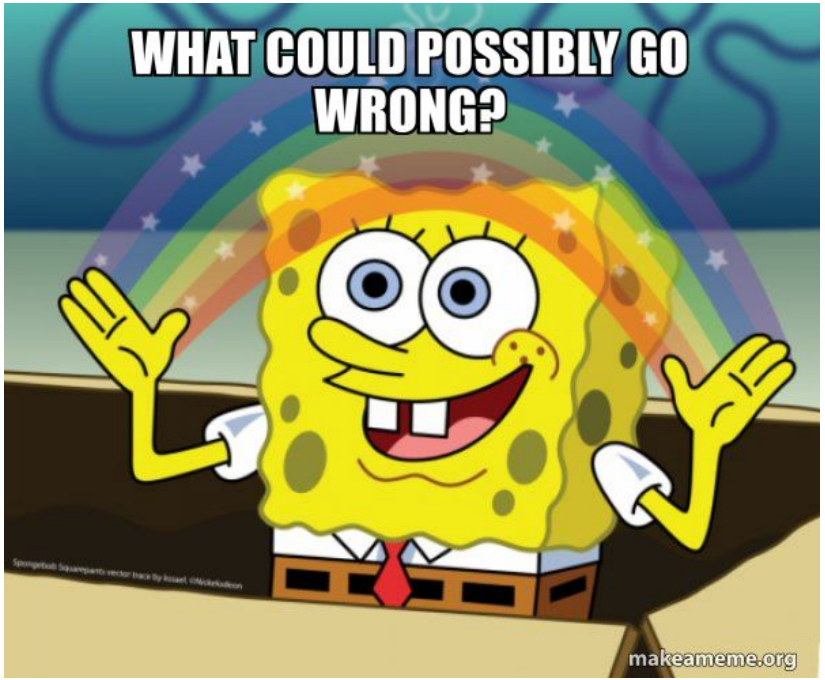
*https://cursor.com/blog/composer-2-5

What If There Is No World?

Extension to the math reasoning

Self-distillation helped agents.

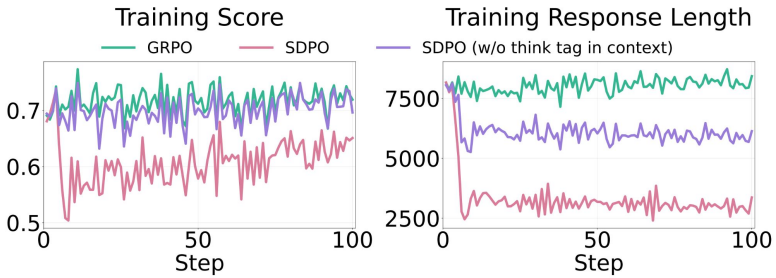
So we tried it on single-turn math reasoning.

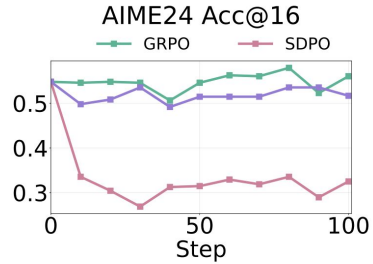


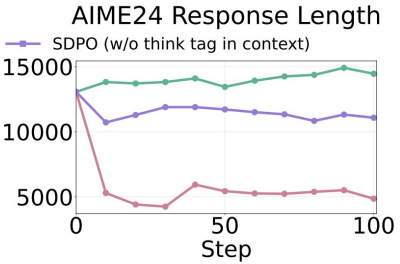
On-Policy Self-Distillation in Math

Purple: Low-information PI
Pink: High-information PI

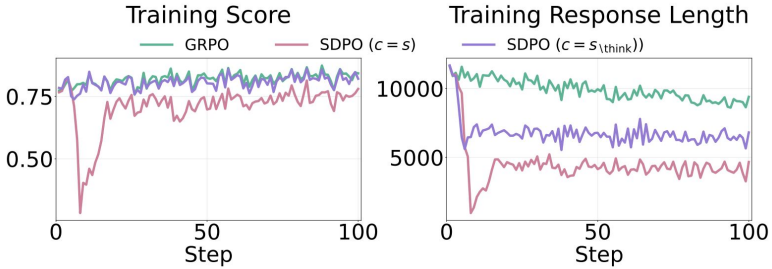
DeepSeek Distilled Qwen 7B

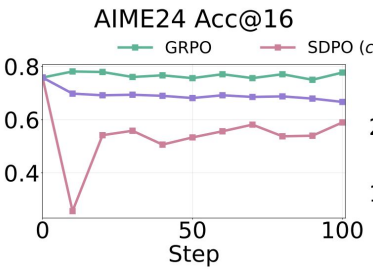


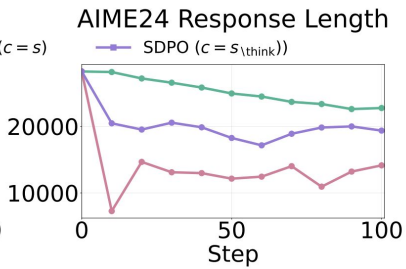




Qwen3 8B







Why Does Self-Distillation (Sometimes) Degrade the Reasoning Capability of LLMs?

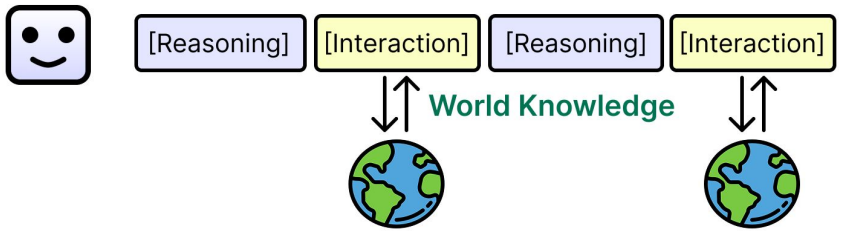
Jeonghye Kim , Xufang Luo , Minbeom Kim , Sangmook Lee , Dohyung Kim, Jiwon Jeon, Dongsheng Li, Yuqing Yang

World-Bayesian vs Self-Bayesian reasoning

Shared Goal of Reasoning: Seeking task-relevant information to reduce task uncertainty.

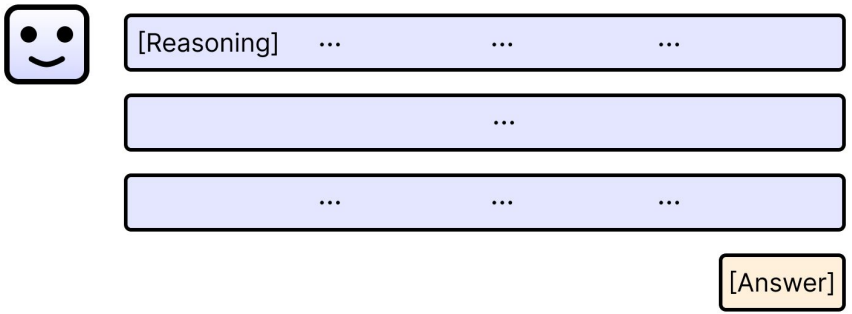
World-Bayesian

The model seeks task-relevant information through interaction with the world.



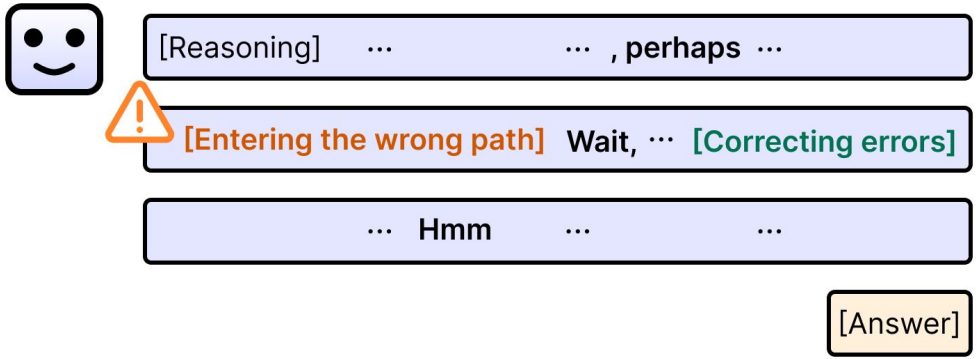
Self-Bayesian

The model seeks task-relevant information by refining its belief based on the problem and its own generated tokens.



Without External Feedback, How Does the Model Know It's Wrong? → Internal Hesitation and Epistemic Doubt

Internal error detection: Epistemic Verbalization



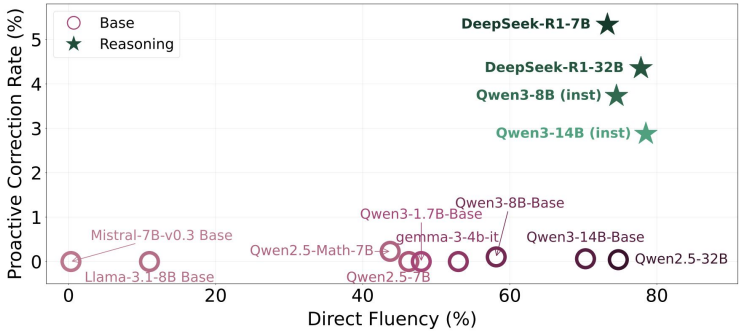
How can the model detect errors when there is no explicit contradiction in its reasoning?

“Wait, is this right?” is not just style. It is an internal error signal.
Epistemic verbalization is where **“proactive”** self-correction begins.

Kim, Jeonghye, et al. "Understanding reasoning in llms through strategic information allocation under uncertainty." *arXiv 2026*.

Internal error detection: Epistemic Verbalization

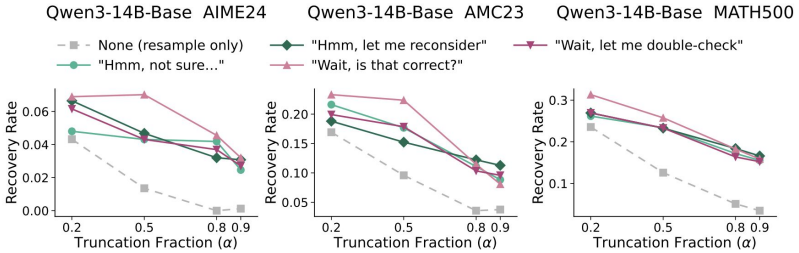
LRMs show stronger proactive correction than standard LLMs, but accuracy is only around 20–30%.



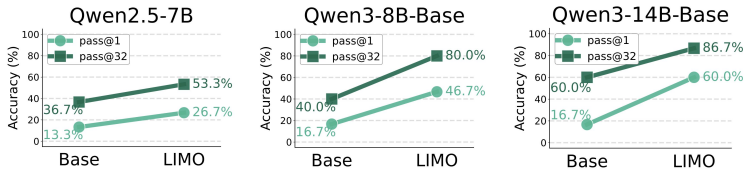
Model	Precision (%)
DeepSeek-R1-Distill-Qwen-7B	37.3
DeepSeek-R1-Distill-Qwen-32B	23.9
Qwen3-8B	20.5
Qwen3-14B	15.8

* Precision of proactive self-correction signals across LRMs

A simple doubt cue can recover failed attempts.



This proactive correction is a linguistic habit,that can be learned quickly with small-scale SFT (only 800 samples).



Kim, Jeonghye, et al. "Understanding reasoning in llms through strategic information allocation under uncertainty." *arXiv 2026*.

LLM Reasoning Behavior Under Richer Information

Prompt for unguided generation	{question} Please reason step by step, and put your final answer within \boxed{}.
Regeneration prompt (followed the prompt in Hübotter et al. (2026))	{question} Please reason step by step, and put your final answer within \boxed{}. Correct solution: {previously correct solution} Correctly solve the original question.

wait	hmm	perhaps	maybe	actually	alternatively	seems	might	likely	check

Table 1: Comparison of response characteristics under varying levels of rich information.

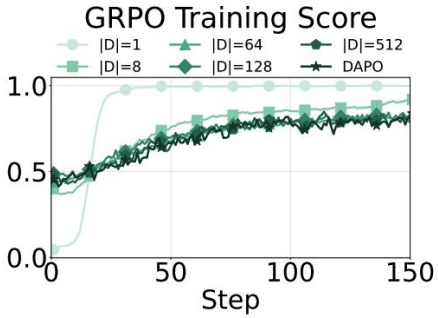
	Avg. Score	Avg. Length	Epistemic Token Count
(1) Unguided	0.30	13,054	182.5
(2) Solution-Guided	0.98	1,873	8.8
(3) Solution-Guided (w/o think tag)	0.78	12,036	159.8
(4) Regeneration-Conditioned	0.95	2,808	24.1

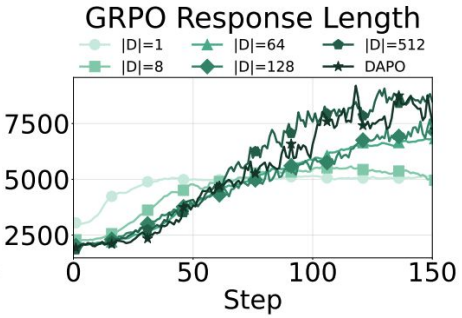
(associated with pink)

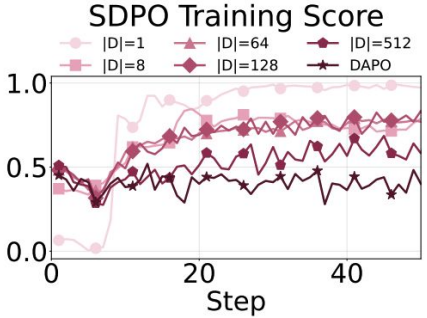
(associated with purple on Slide 20)

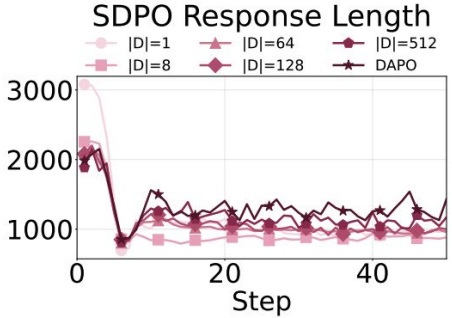
Self-distillation can erase the internal error signal

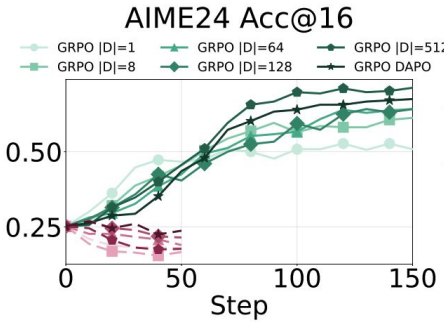
Task Coverage and Generalization Ability

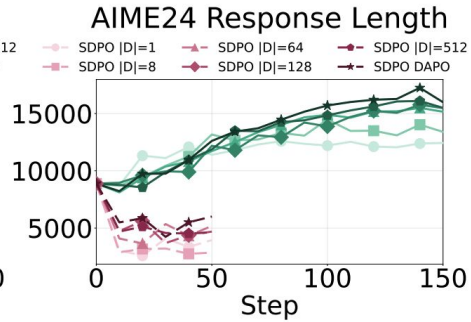










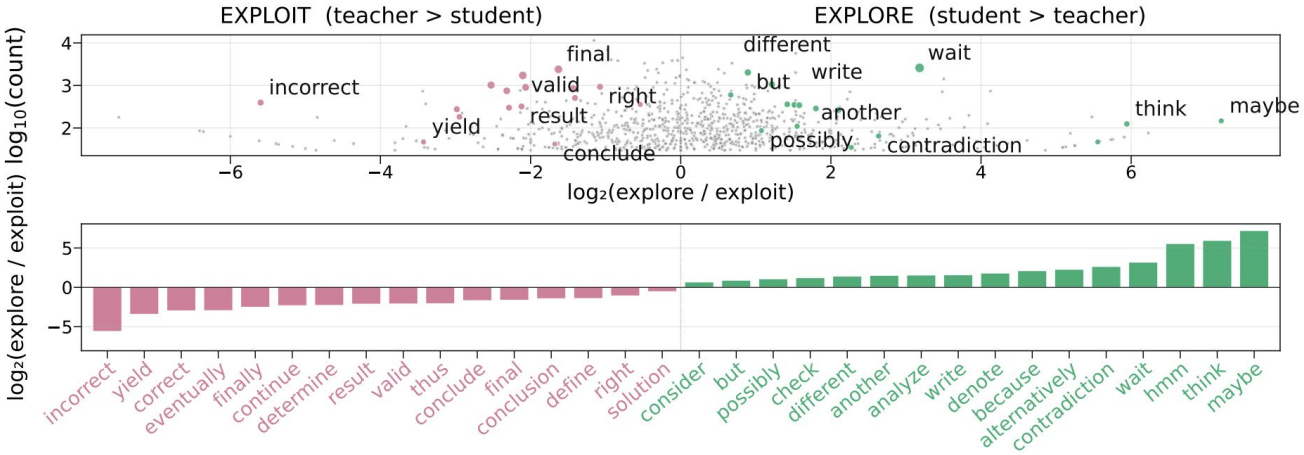


If the PI-conditioned teacher removes uncertainty, should the student still blindly imitate it?

Rebellious Student: Reversing Teacher Signals for Reasoning Exploration with Self-Distilled RLVR

Jeonghye Kim*, **Jiwon Jeon***, Dongsheng Li, Yuqing Yang

Reading the Teacher Signal in Reverse



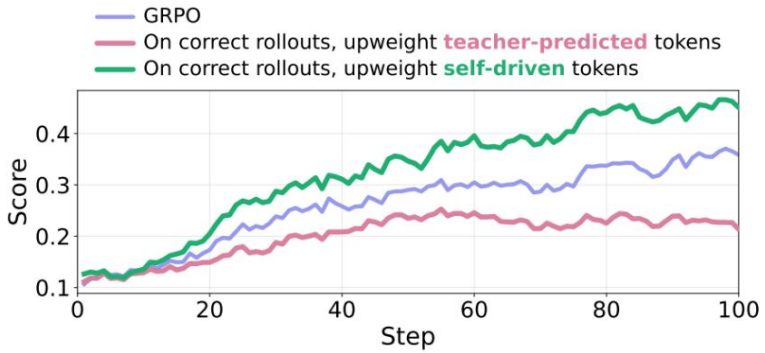
But on correct rollouts, teacher-student disagreement can mean something different.

Teacher > Student
Tokens the teacher would have predicted → exploit direction

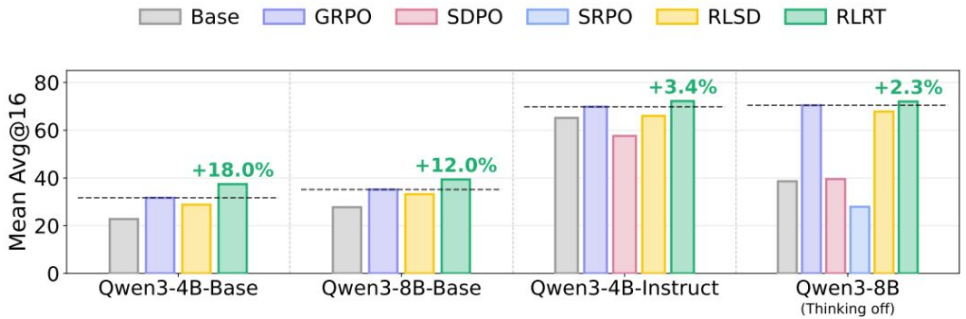
Student > Teacher
Tokens the student chose against the teacher, but still succeeded → explore direction

Reversal: instead of suppressing these student-driven tokens, reinforce them.

Rebellious Student: Results



(a) Training reward on Qwen3-4B-Base by token weighting



(b) Mean benchmark score across models and methods

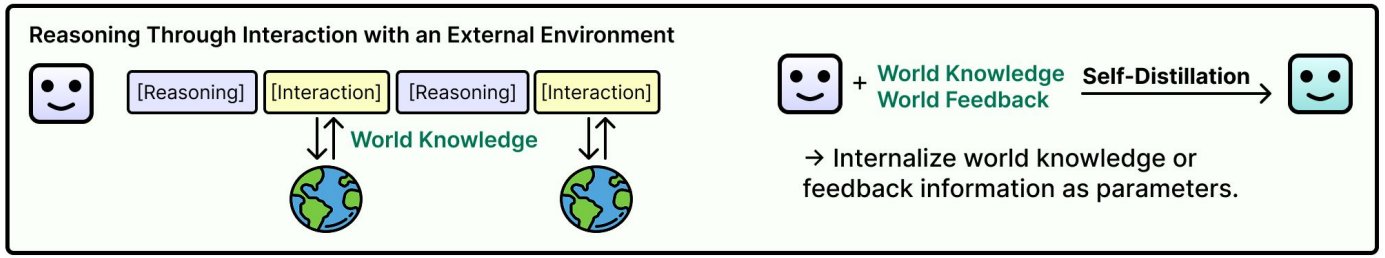
On correct rollouts:

Standard self-distillation
Upweight teacher-predicted tokens
→ pulls the student back to the teacher's path

RLRT
Upweight self-driven tokens
→ reinforces successful deviations from the teacher

Summary

World-Baysian Reasoning (e.g. long horizon agent tasks)



Self-Baysian Reasoning (e.g. math)