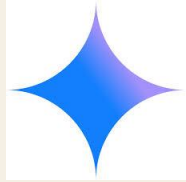


Continual Adaptation of Language Models Through Memory and Finetuning

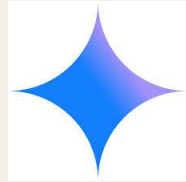
Yuwei Zhang
Halicioğlu Data Science Institute
UC San Diego

Why LLMs should continually adapt?



- Pretraining/post-training gives a model general capability

Why LLMs should continually adapt?



- Pretraining/post-training gives a model general capability
- But the models are static during deployment which will cause consistent failures

Why LLMs should continually adapt?





- Pretraining/post-training gives a model general capability
- But the models are static during deployment which will cause consistent failures

LLMs should continually adapt to the world

Personalization





Without memory	With memory
<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>	<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>
<p>Great add-on city for a work trip: Singapore is compact, safe, excellent for solo exploring, and easy to do in 2-4 free days. July is hot and humid; expect daytime highs around 31-33°C / 88-91°F, warm nights, and sudden rain, so plan outdoor time early morning or after 5 p.m. and keep midday for museums, malls, cafés, or hotel downtime. Singapore's official weather service notes typical daily highs of 31-33°C and lows of 23-25°C. (Singapore Weather Service)</p> 	<p>I'll build this around your known constraints: a few free days after work, hot/humid July weather, wildlife/photography/nature bias, strong AC needs for sleep, and good but not overly social dining. I'll verify current July logistics and then give a practical short itinerary rather than a generic Singapore list.</p> 
<p>The model produces a generic response that is more touristy, doesn't help with hotel booking, and largely ignores the user's interests.</p>	<p>The model produces a response that is personalized to the user's interests in wildlife photography, quiet dinners, and their priorities when booking a hotel.</p>

[Chatgpt's Dreaming Memory Blog](#)

LLMs should continually adapt to the world

Personalization



Without memory	With memory
<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>	<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>
<p>Great add-on city for a work trip: Singapore is compact, safe, excellent for solo exploring, and easy to do in 2-4 free days. July is hot and humid: expect daytime highs around 31-33°C / 88-91°F, warm nights, and sudden rain, so plan outdoor time early morning or after 5 p.m. and keep midday for museums, malls, cafés, or hotel downtime. Singapore's official weather service notes typical daily highs of 31-33°C and lows of 23-25°C. (Singapore Weather Service)</p> 	<p>I'll build this around your known constraints: a few free days after work, hot/humid July weather, wildlife/photography/nature bias, strong AC needs for sleep, and good but not overly social dining. I'll verify current July logistics and then give a practical short itinerary rather than a generic Singapore list.</p> 
<p><i>The model produces a generic response that is more touristy, doesn't help with hotel booking, and largely ignores the user's interests.</i></p>	<p><i>The model produces a response that is personalized to the user's interests in wildlife photography, quiet dinners, and their priorities when booking a hotel.</i></p>

[Chatgpt's Dreaming Memory Blog](#)

New knowledge

Put together a plan for the v2.3 release. Summarize what's ready, what's blocked, and what needs follow-up

+ Company knowledge 3

Thought for 51s

- GitHub v2.3 Release
- Linear Pending QA
- Slack #mobile-release
- 18 source

I compiled a release plan for the **v2.3 mobile update** by scanning our GitHub repo for open TODO checking Linear for related tickets, and reviewing Slack engineering threads for unresolved bug reports. The summary below outlines what's already tracked, what's outstanding, and what still needs to be logged for the upcoming release. v2.3 Release



[Chatgpt's Company Knowledge](#)

LLMs should continually adapt to the world

Personalization

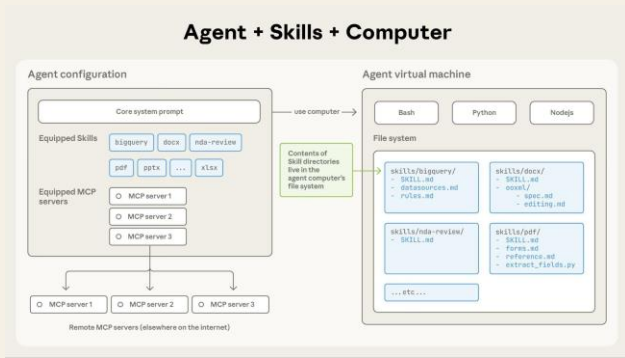
Improved capability



Without memory	With memory
<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>	<p>Help me plan a trip to Singapore. I'm going to be there for a work trip in July and will have a few days free at the end.</p>
<p>Great add-on city for a work trip: Singapore is compact, safe, excellent for solo exploring, and easy to do in 2-4 free days. July is hot and humid; expect daytime highs around 31-33°C / 88-91°F, warm nights, and sudden rain, so plan outdoor time early morning or after 5 p.m. and keep midday for museums, malls, cafes, or hotel downtime. Singapore's official weather service notes typical daily highs of 31-33°C and lows of 23-25°C (Singapore Weather Service)</p> 	<p>I'll build this around your known constraints: a few free days after work, hot/humid July weather, wildlife/photography/nature bias, strong AC needs for sleep, and good but not overly social dining. I'll verify current July logistics and then give a practical short itinerary rather than a generic Singapore list.</p> 
<p>The model produces a generic response that is more touristy, doesn't help with hotel booking, and largely ignores the user's interests.</p>	<p>The model produces a response that is personalized to the user's interests in wildlife photography, quiet dinners, and their priorities when booking a hotel.</p>

[Chatgpt's Dreaming Memory Blog](#)

New knowledge



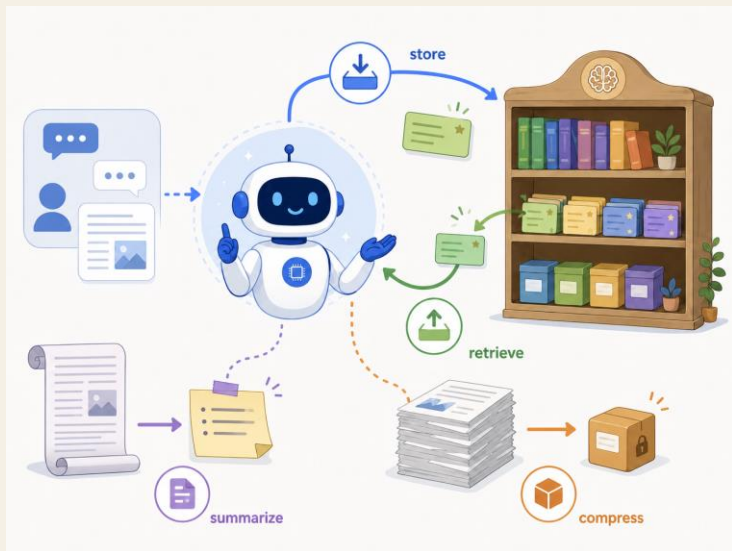
[Anthropic's Agent Skills Doc](#)

The screenshot shows a ChatGPT conversation where the user asks for a release plan for v2.3. The AI response is a detailed plan for the v2.3 mobile update, mentioning scanning GitHub repos, checking Linear for tickets, and reviewing Slack threads for bug reports. The response is structured with a title 'Thought for 51s' and includes a table of sources like GitHub, Linear, and Slack.

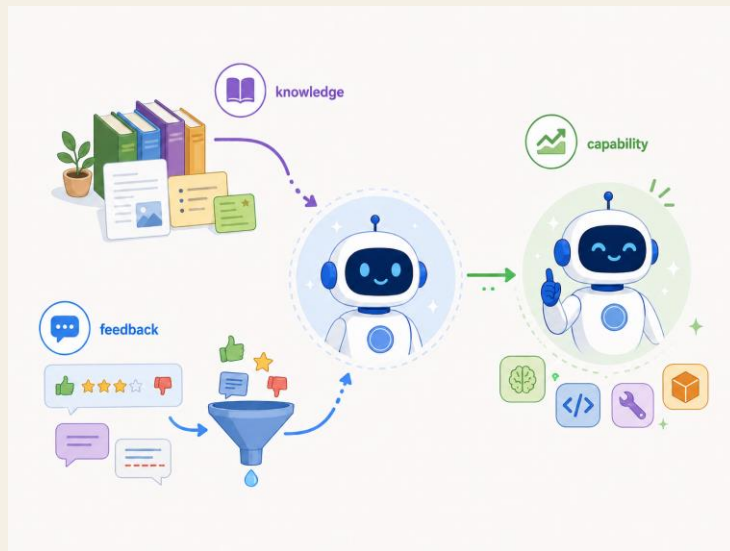
[Chatgpt's Company Knowledge](#)

How should LLMs continually adapt?

Memory



Finetuning

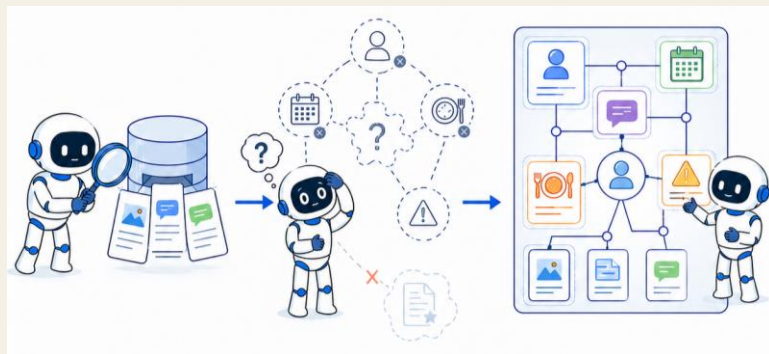


Part 1: Memory



[LongMemEval ICLR'25]: long-term conversation capability
[ImplexConv EMNLP'25]: Implicit personalization through multi-session conversations
[CoMem ICML'26]: efficient context management

Maybe memory is just RAG



- Relevant evidence can be implicit [Li et al. EMNLP'25]
- Memory needs complex organization and structuring [Xu et al. NeurIPS'25]

Challenge 1: Implicit Evidence Retrieval is Hard

Aug 15, 2024
I love playing sports like basketball and swimming. I'm going on a sports stadium tour!

Sep 15, 2024
I broke my leg in a car accident...

Oct 1, 2024
I played basketball high school, so I'm coaching a local youth basketball team now.

Oct 13, 2024
I'm learning about the science behind athletic performance

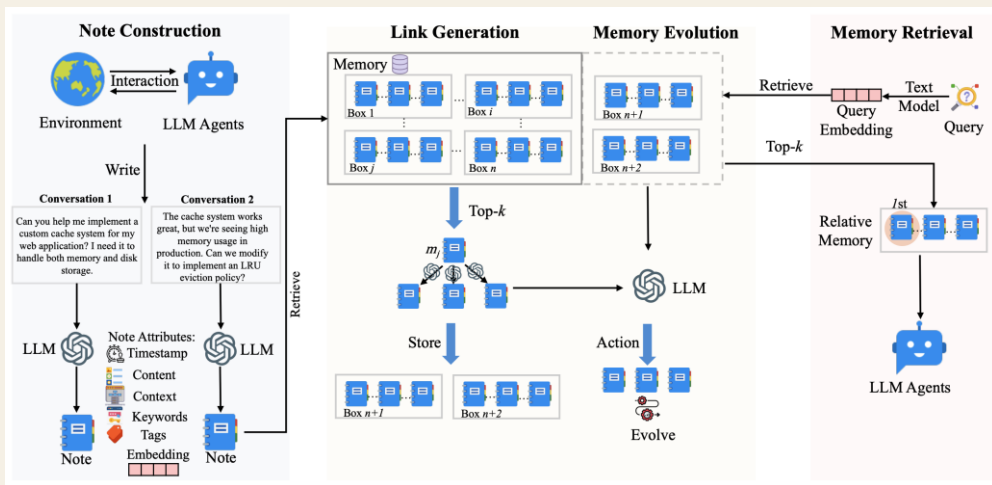
Oct 31, 2024
Starting a sports blog to share my journey!!

Nov 15, 2024
I'm bored. Can you give me some suggestions?

Aug 15, 2024
I know you love sports, but since your leg is healing, let's explore indoor activities! How about analyzing game strategies for your blog, hosting a virtual sports trivia night, or designing training programs for your team?

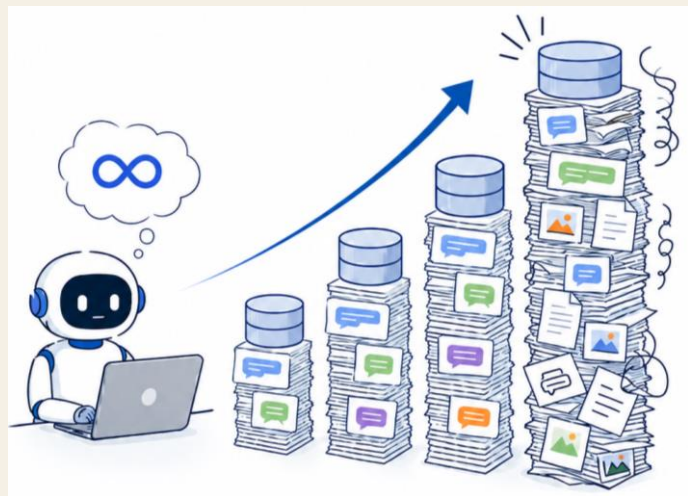
- Retrieval may focus on all sports-related memories
- Actual useful memory is more subtle.
- Search should narrow by reasoning not similarity alone

Challenge 2: Generalizable Memory Design



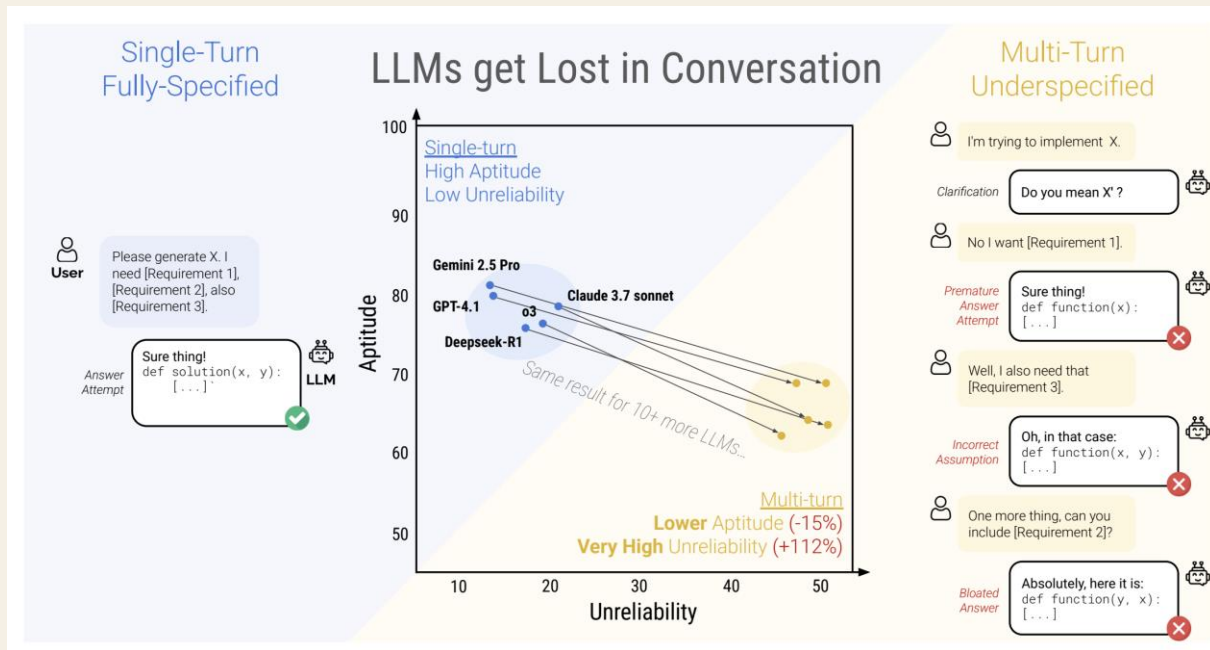
- Not just append transcript chunks
- Memories need structure and relations
- New memories may change old context
- Design should transfer across systems

Maybe memory is just long-context



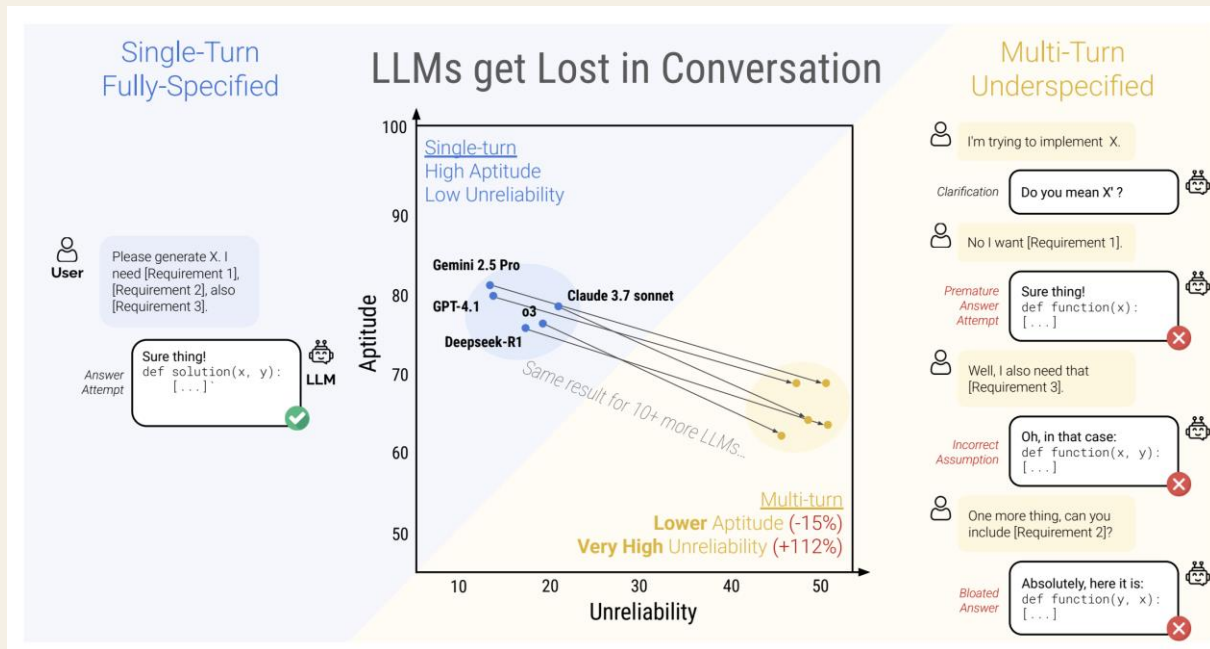
- Multi-turn interaction is unreliable [Laban et al. ICLR'25]
- Long histories contain distracting noise [Wu et al. ICLR'25]
- Linear growth of context length increases the response latency

Challenge 1: Multi-turn Interaction is More Unreliable



- Same task information
- Different interaction format
- Multi-turn setting drops performance

Challenge 1: Multi-turn Interaction is More Unreliable



- Same task information
- Different interaction format
- Multi-turn setting drops performance

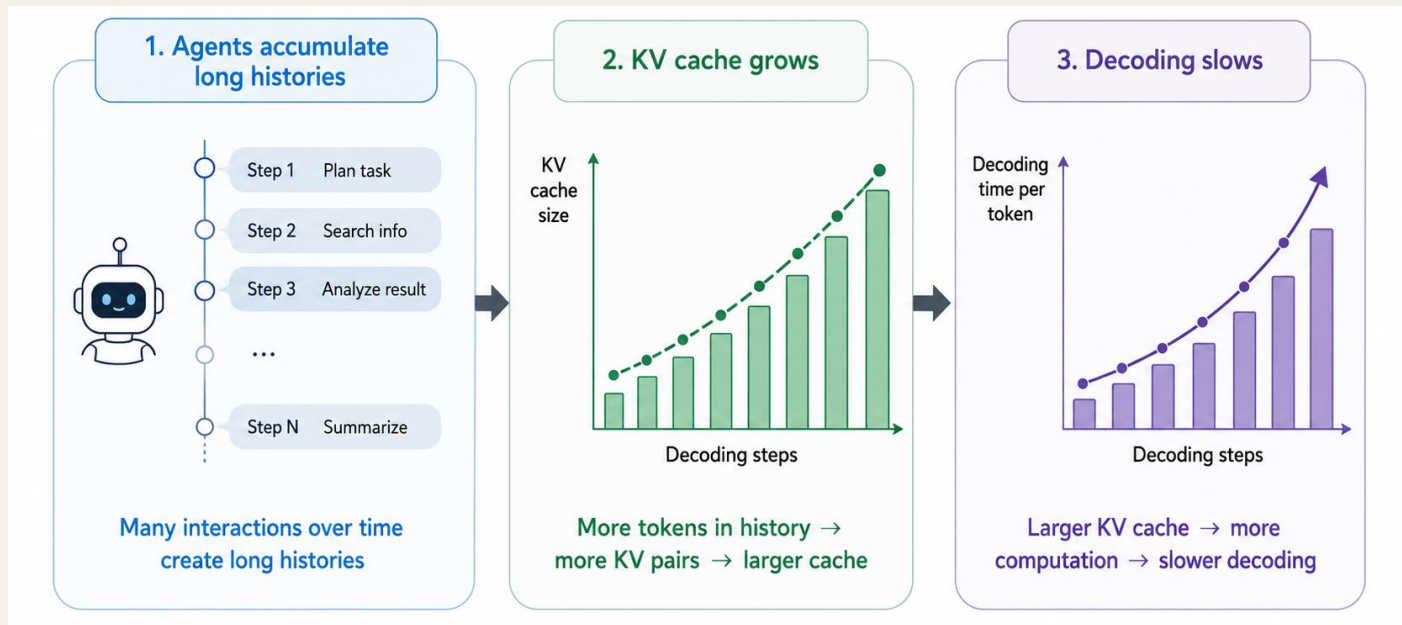
Having the history in context does not guarantee reliable use of it.

Challenge 2: Long Histories Contain Distracting Noise

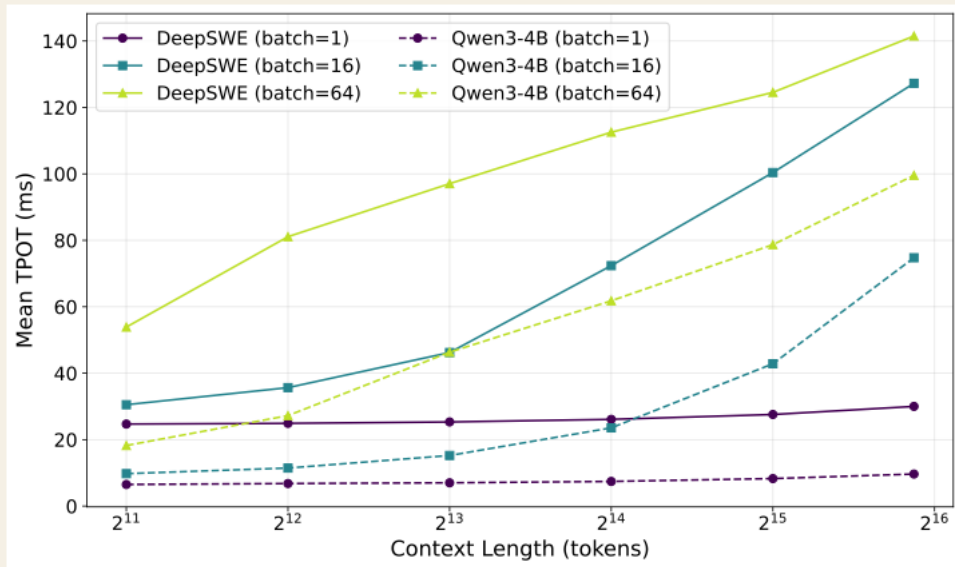
Model	Size	Oracle	S	% Drop
No Chain-of-Note				
GPT-4o	-	0.870	0.606	30.3%↓
Llama 3.1 Instruct	70B	0.744	0.334	55.1%↓
Llama 3.1 Instruct	8B	0.710	0.454	36.1%↓
Phi-3 128k Instruct	14B	0.702	0.380	45.9%↓
Phi-3.5 Mini Instruct	4B	0.660	0.342	48.1%↓
With Chain-of-Note				
GPT-4o	-	0.924	0.640	30.7%↓
Llama 3.1 Instruct	70B	0.848	0.286	66.3%↓
Llama 3.1 Instruct	8B	0.710	0.420	40.8%↓
Phi-3 128k Instruct	14B	0.722	0.344	52.4%↓
Phi-3.5 Mini Instruct	4B	0.652	0.324	50.3%↓

Long-context models still struggle when the answer is buried in long, realistic interaction histories.

Challenge 3: Conversation Growth Creates Efficiency Cost

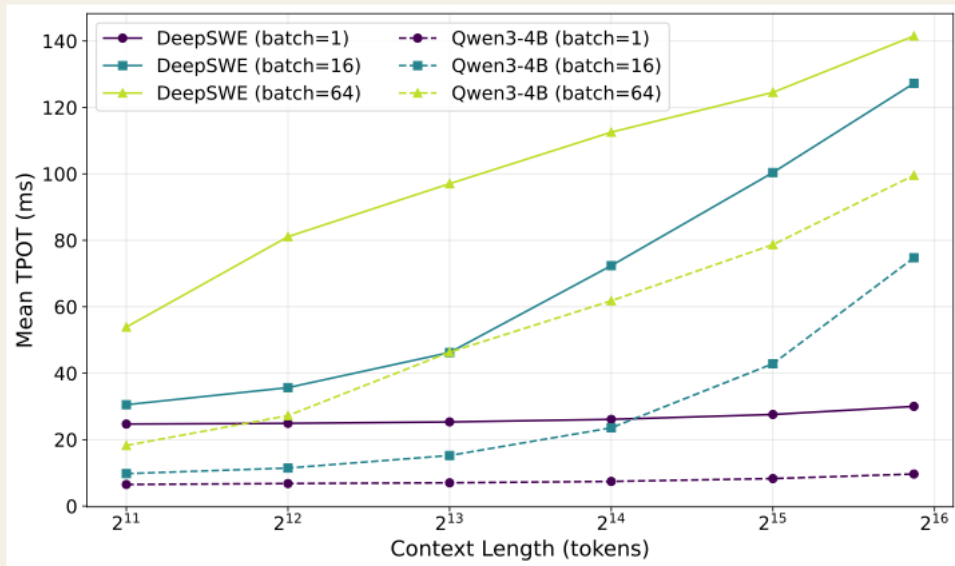


Challenge 3: Conversation Growth Creates Efficiency Cost



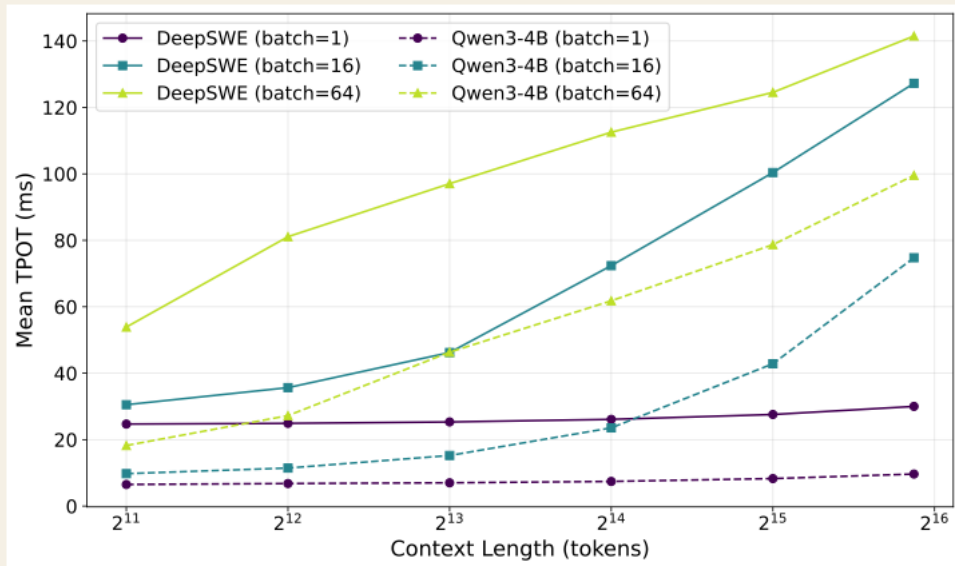
- Batch=1: TPOT is flat

Challenge 3: Conversation Growth Creates Efficiency Cost



- Batch=1: TPOT is flat
- Batch=16/64: TPOT scales linearly with context length (memory-bound)

Challenge 3: Conversation Growth Creates Efficiency Cost



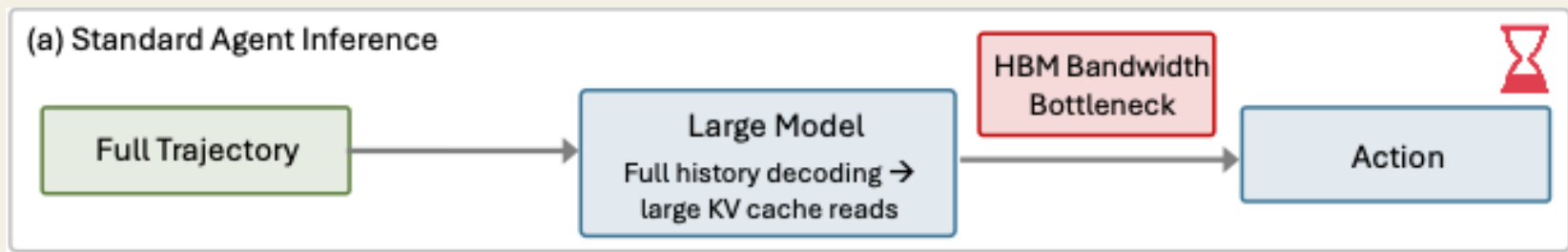
- Batch=1: TPOT is flat
- Batch=16/64: TPOT scales linearly with context length (memory-bound)
- Larger model (32B) suffers more than smaller model (4B) at the same context length

CoMem: Context Management with A Decoupled Long-Context Model

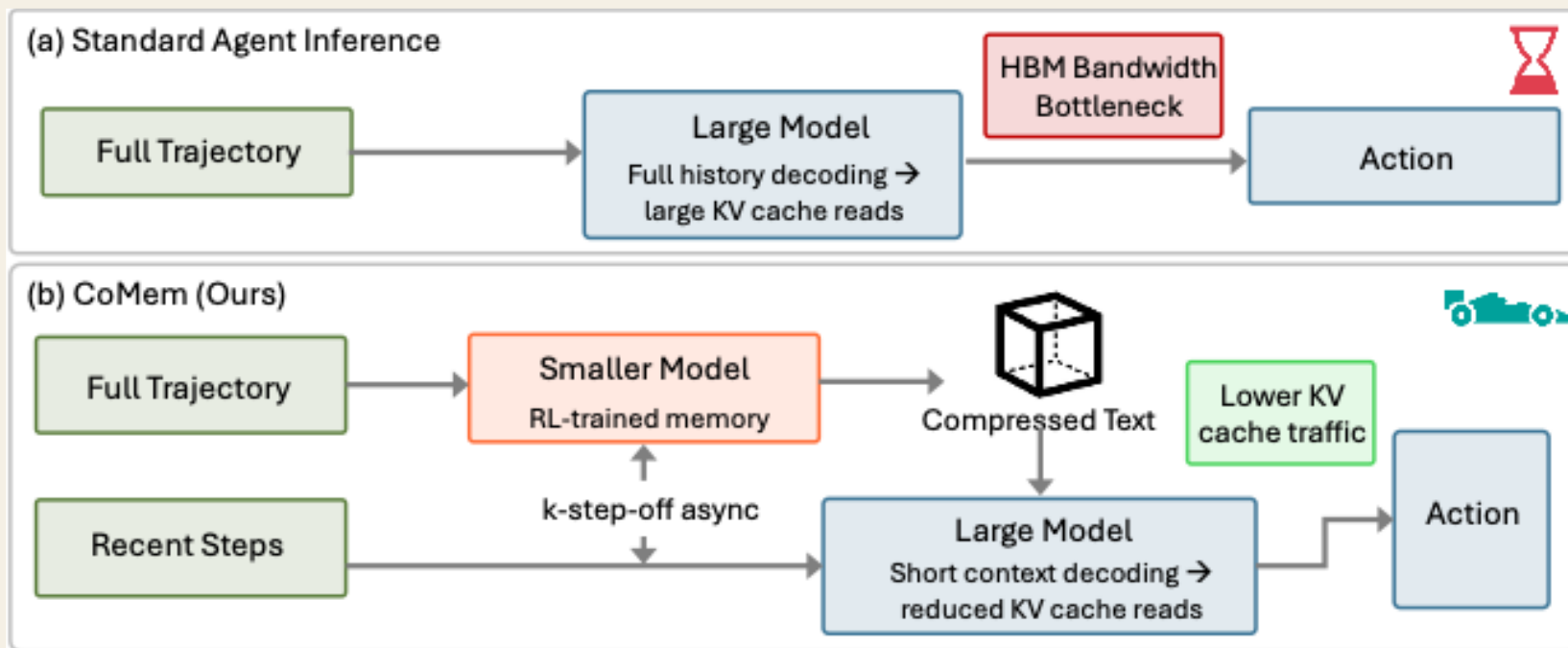
Yuwei Zhang, Chengyu Dong, Shuwei Jin, Changlong Yu,
Hejie Cui, Hongye Jin, Xinyang Zhang, Hamed Bonab, Colin Lockard,
Jianshu Chen, Zhenyu Shi, Jingbo Shang, Xian Li, Bing Yin



CoMem Overview

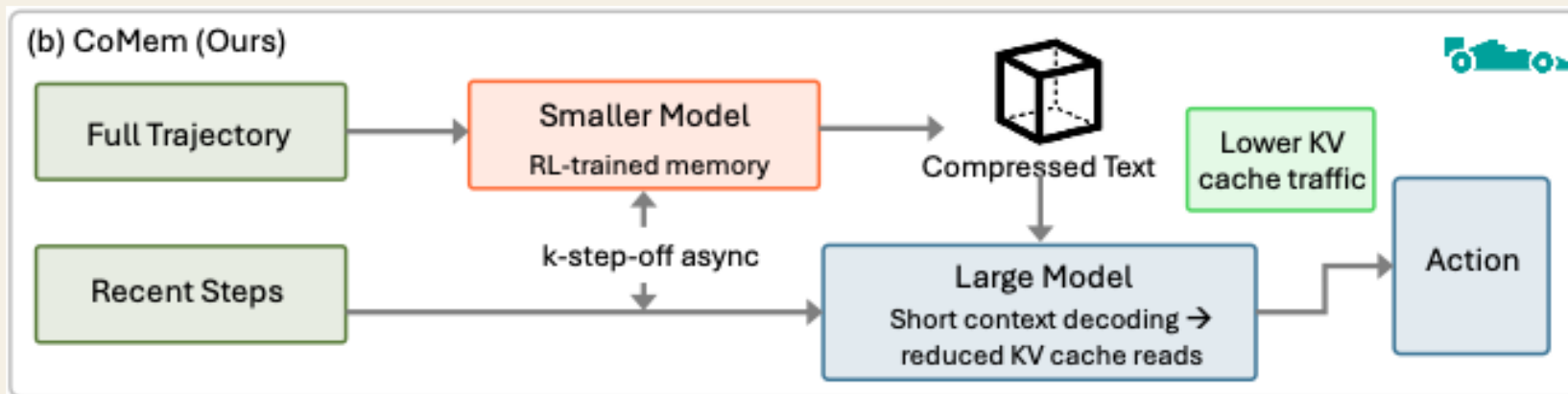


CoMem Overview



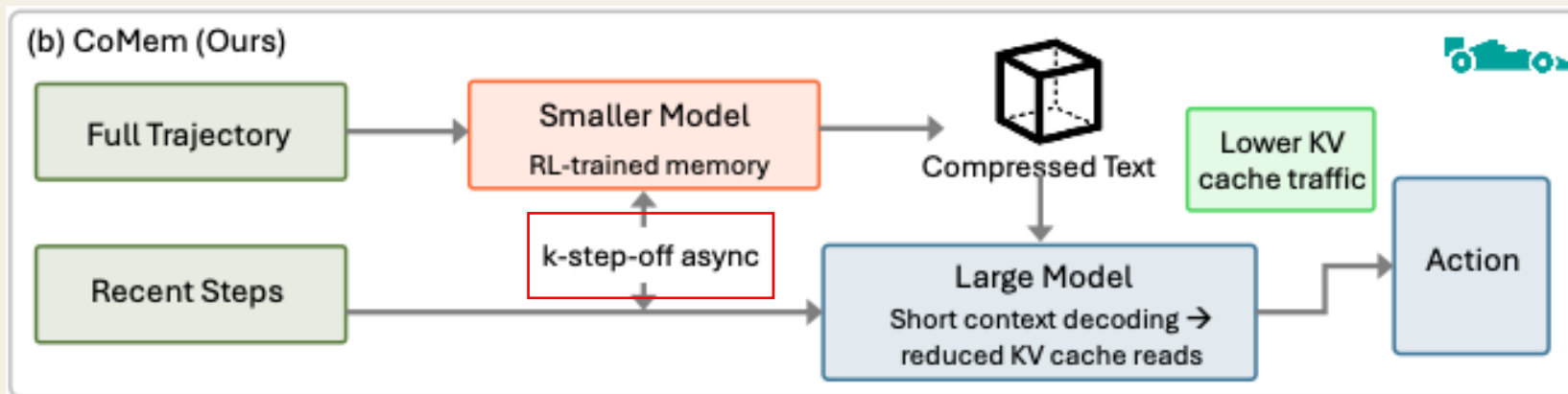
CoMem Overview

TL;DR: CoMem moves long-history processing out of the main agent path, it avoids the latency wall of full-context inference.

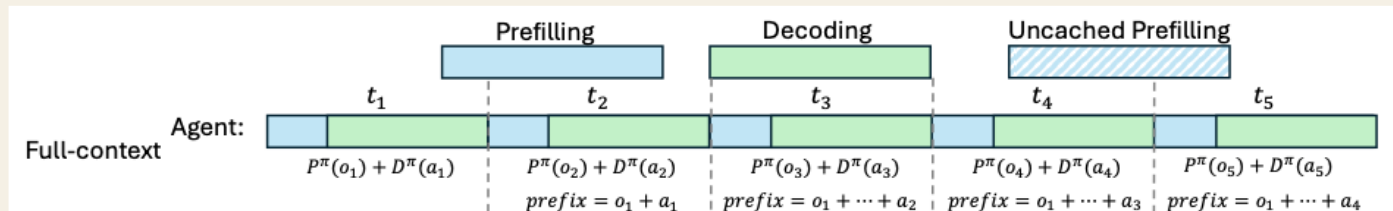


CoMem Overview

TL;DR: CoMem moves long-history processing out of the main agent path, it avoids the latency wall of full-context inference.

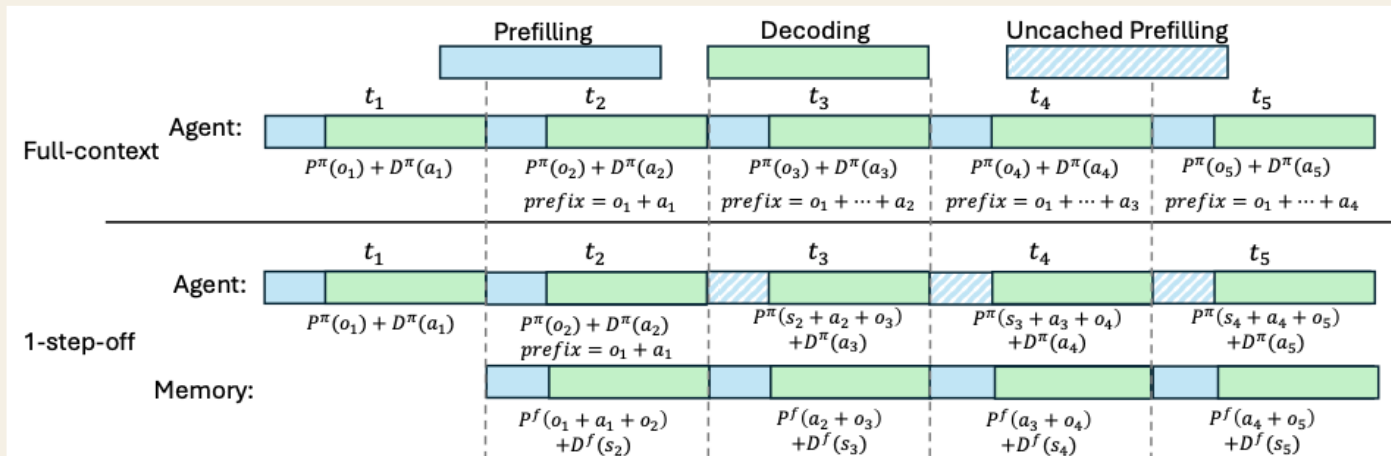


K-step-off pipeline



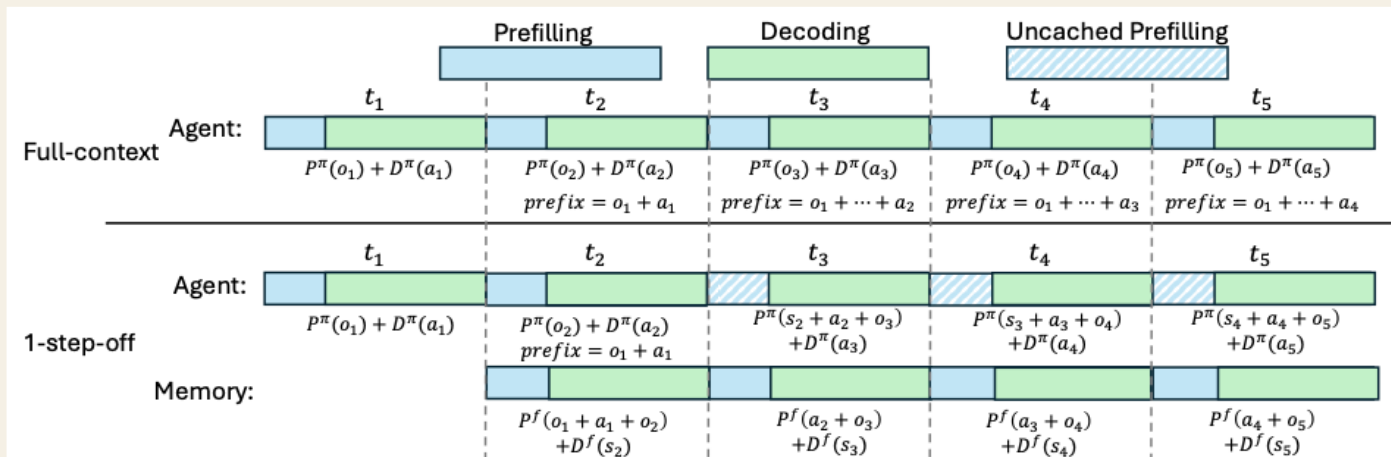
o_i - Observation
 a_i - Action
 s_i - Summary
 P^π - Agent Prefill
 D^π - Agent Decode
 P^f - Memory Prefill
 D^f - Memory Decode

K-step-off pipeline



o_i - Observation
 a_i - Action
 s_i - Summary
 P^π - Agent Prefill
 D^π - Agent Decode
 P^f - Memory Prefill
 D^f - Memory Decode

K-step-off pipeline



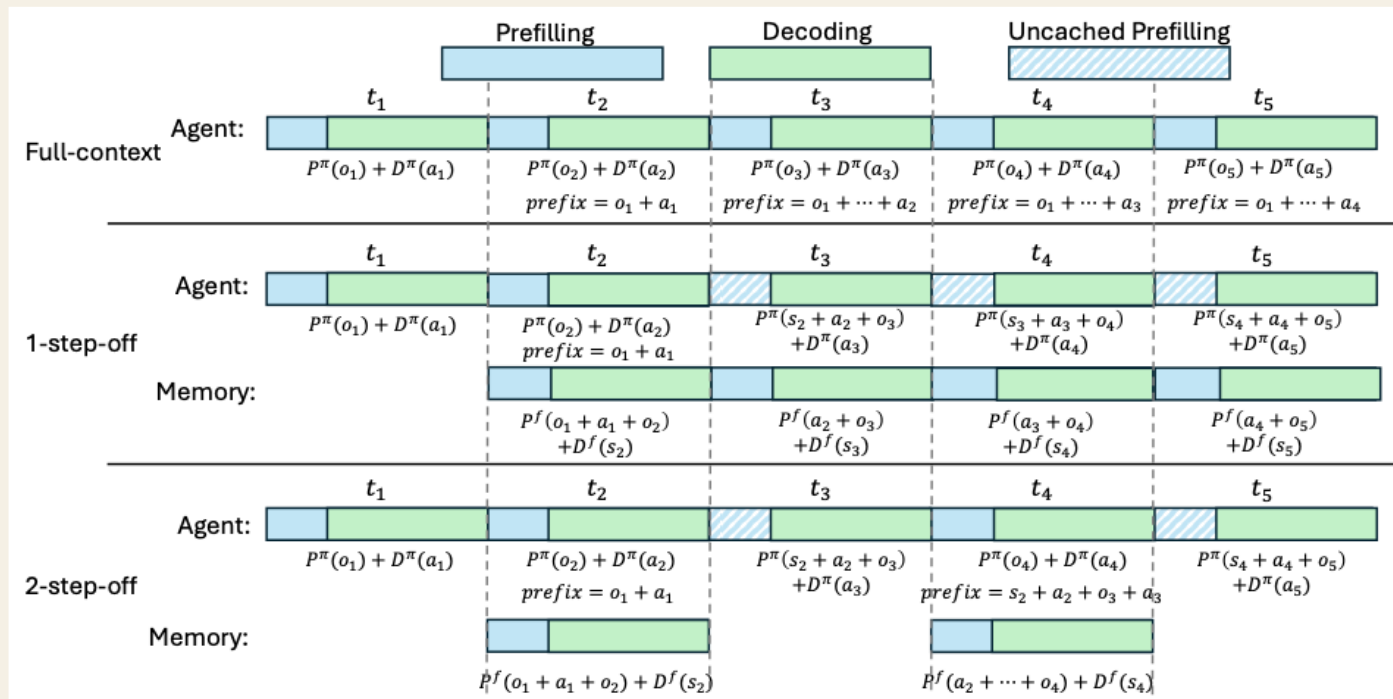
o_i - Observation
 a_i - Action
 s_i - Summary
 P^π - Agent Prefill
 D^π - Agent Decode
 P^f - Memory Prefill
 D^f - Memory Decode

$$\frac{L_{\text{sum}}}{L_{\text{full}}} < \frac{1}{1 + \frac{W}{Y \cdot S \cdot P}}$$

On a single A100 80G, for a 32K prefix, summary needs to be < 7.36K tokens (23% compression ratio) to guarantee net speedup.

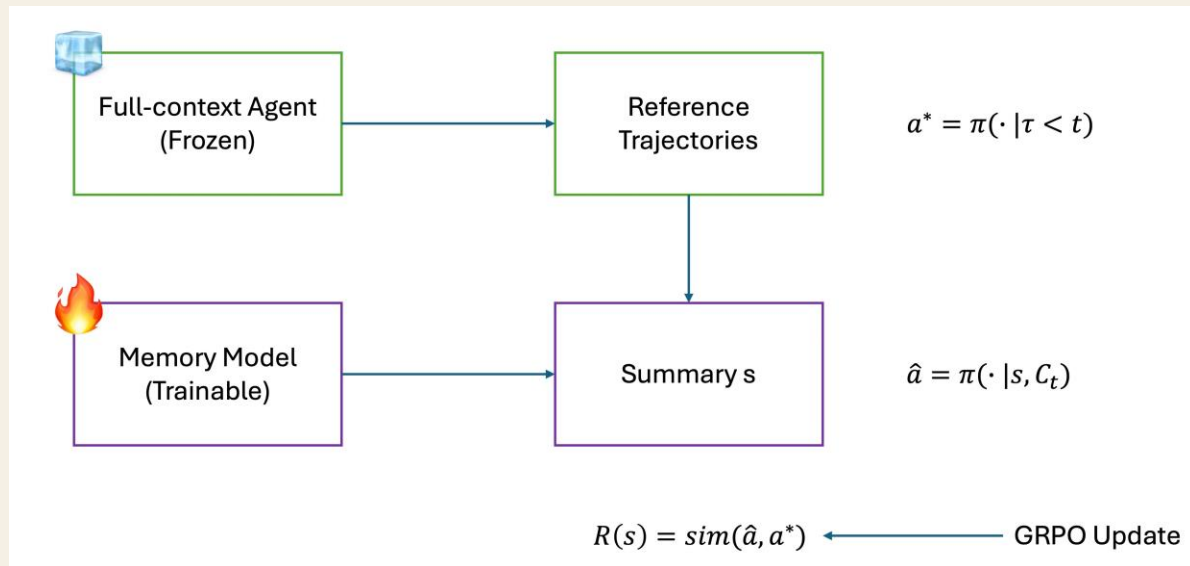
✓ Easily achievable with summarization.

K-step-off pipeline

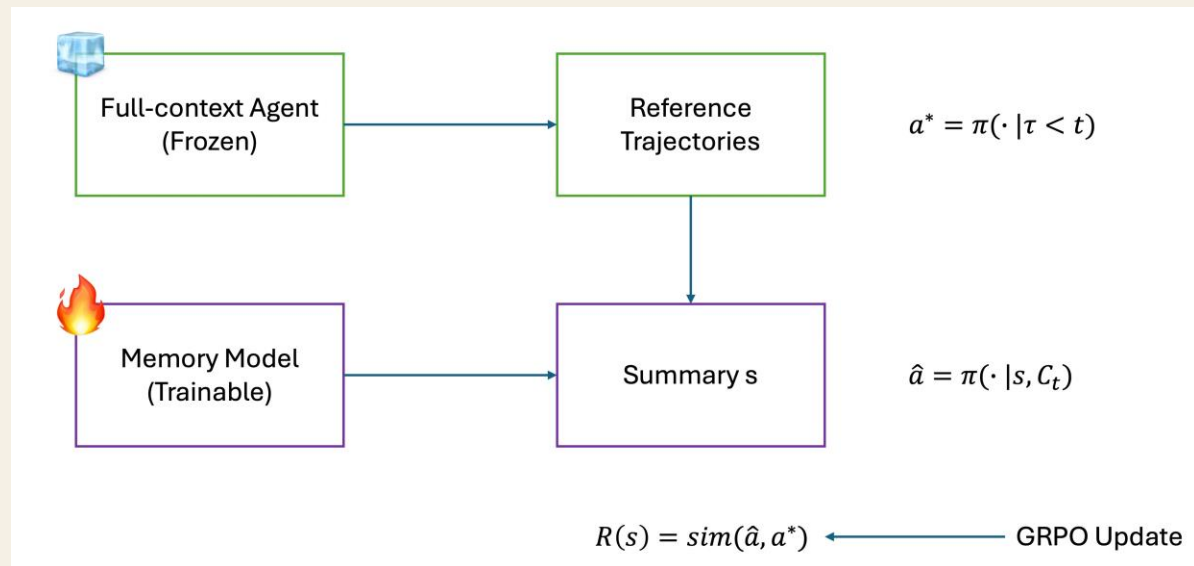


o_i - Observation
 a_i - Action
 s_i - Summary
 P^π - Agent Prefill
 D^π - Agent Decode
 P^f - Memory Prefill
 D^f - Memory Decode

Summary Model Training



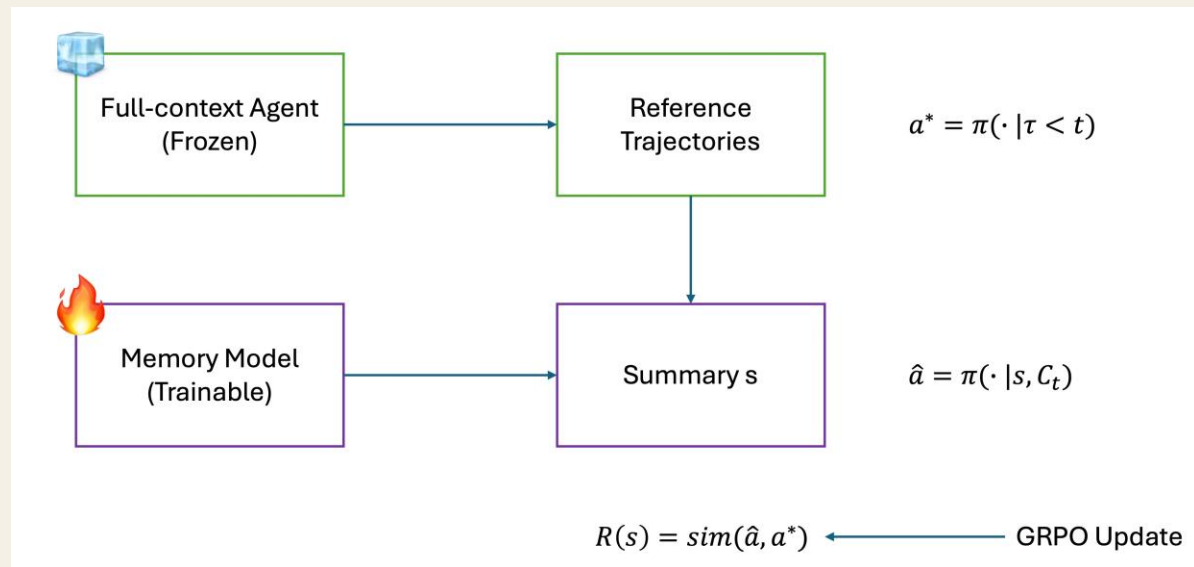
Summary Model Training



Similarity function measures how close two proposed actions are:

- same function called
- same arguments employed
- same file path employed
- edit distance between strings
- ...

Summary Model Training



Such a training method prevents sparse rewards and long-horizon training, with only offline interactions.

Main Results

Table 1. Main results. For latency, we show the total inference time for 128 issues.

Agent	Memory	%Resolved	#Tool Calls	w/o CPU Offload		w/ CPU Offload	
				Latency ($\times 128/s$)	Speedup	Latency ($\times 128/s$)	Speedup
DeepSWE (32B Dense)	Full-Context	40.4	34.39	9457.78	1 \times	6390.62	1 \times
	Qwen3-4B (base)	29.9	24.74	4360.72	2.17 \times	3649.61	1.75 \times
	Qwen3-4B (SFT)	39.8	35.61	5232.64	1.81 \times	4168.85	1.53 \times
	GRPO _{AC}	41.0	40.69	5622.33	1.68\times	4400.89	1.45 \times
Qwen3-Coder-Max (480B A35B)	Full-Context	57.2	36.68	6291.11	1 \times	5129.90	1 \times
	No Summary	46.6	25.20	3780.74	1.66 \times	3421.78	1.50 \times
	Qwen3-4B (base)	42.2	39.40	3421.63	1.84 \times	3138.97	1.63 \times
	Qwen3-4B (SFT)	47.3	43.39	3819.70	1.65 \times	3337.58	1.54 \times
	GRPO _{AC}	51.0	44.55	3917.68	1.61 \times	3594.51	1.43 \times
GLM-4.7 (355B 32B)	Full-Context	69.0	55.11	6361.49	1 \times	5869.42	1 \times
	No Summary	59.4	46.82	4216.83	1.51 \times	3842.17	1.53 \times
	Qwen3-4B (base)	58.3	49.32	3928.44	1.62 \times	3526.91	1.66 \times
	Qwen3-4B (SFT)	61.3	52.74	4087.92	1.56 \times	3668.53	1.60 \times
	GRPO _{AC}	62.7	51.21	3318.42	1.92 \times	2821.38	2.08 \times

For mid-sized model, we get 1.7x speedup for free.

Main Results

Table 1. Main results. For latency, we show the total inference time for 128 issues.

Agent	Memory	%Resolved	#Tool Calls	w/o CPU Offload		w/ CPU Offload	
				Latency ($\times 128/s$)	Speedup	Latency ($\times 128/s$)	Speedup
DeepSWE (32B Dense)	Full-Context	40.4	34.39	9457.78	1 \times	6390.62	1 \times
	Qwen3-4B (base)	29.9	24.74	4360.72	2.17 \times	3649.61	1.75 \times
	Qwen3-4B (SFT)	39.8	35.61	5232.64	1.81 \times	4168.85	1.53 \times
	GRPO _{AC}	41.0	40.69	5622.33	1.68 \times	4400.89	1.45 \times
Qwen3-Coder-Max (480B A35B)	Full-Context	57.2	36.68	6291.11	1 \times	5129.90	1 \times
	No Summary	46.6	25.20	3780.74	1.66 \times	3421.78	1.50 \times
	Qwen3-4B (base)	42.2	39.40	3421.63	1.84 \times	3138.97	1.63 \times
	Qwen3-4B (SFT)	47.3	43.39	3819.70	1.65 \times	3337.58	1.54 \times
	GRPO _{AC}	51.0	44.55	3917.68	1.61 \times	3594.51	1.43 \times
GLM-4.7 (355B 32B)	Full-Context	69.0	55.11	6361.49	1 \times	5869.42	1 \times
	No Summary	59.4	46.82	4216.83	1.51 \times	3842.17	1.53 \times
	Qwen3-4B (base)	58.3	49.32	3928.44	1.62 \times	3526.91	1.66 \times
	Qwen3-4B (SFT)	61.3	52.74	4087.92	1.56 \times	3668.53	1.60 \times
	GRPO _{AC}	62.7	51.21	3318.42	1.92 \times	2821.38	2.08\times

For strongest agent, GRPO recovers to 62.7% while delivering 2x speedup with or without CPU offload.

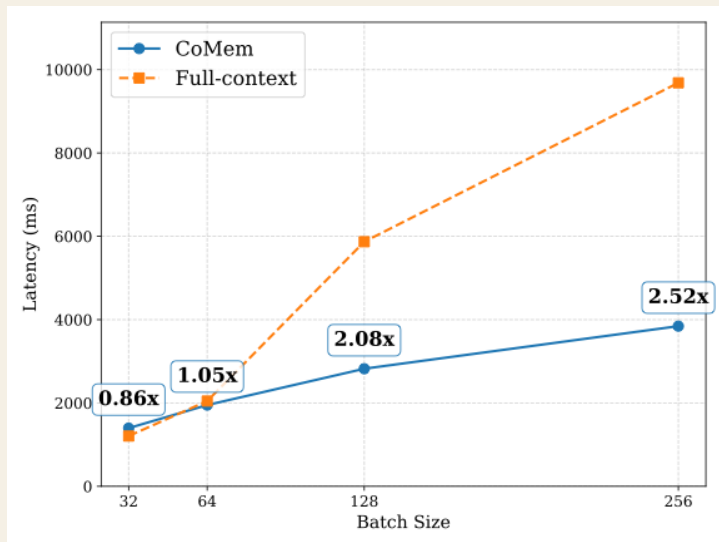
Main Results

Table 1. Main results. For latency, we show the total inference time for 128 issues.

Agent	Memory	%Resolved	#Tool Calls	w/o CPU Offload		w/ CPU Offload	
				Latency ($\times 128/s$)	Speedup	Latency ($\times 128/s$)	Speedup
DeepSWE (32B Dense)	Full-Context	40.4	34.39	9457.78	1 \times	6390.62	1 \times
	Qwen3-4B (base)	29.9	24.74	4360.72	2.17 \times	3649.61	1.75 \times
	Qwen3-4B (SFT)	39.8	35.61	5232.64	1.81 \times	4168.85	1.53 \times
	GRPO _{AC}	41.0	40.69	5622.33	1.68 \times	4400.89	1.45 \times
Qwen3-Coder-Max (480B A35B)	Full-Context	57.2	36.68	6291.11	1 \times	5129.90	1 \times
	No Summary	46.6	25.20	3780.74	1.66 \times	3421.78	1.50 \times
	Qwen3-4B (base)	42.2	39.40	3421.63	1.84 \times	3138.97	1.63 \times
	Qwen3-4B (SFT)	47.3	43.39	3819.70	1.65 \times	3337.58	1.54 \times
	GRPO _{AC}	51.0	44.55	3917.68	1.61 \times	3594.51	1.43 \times
GLM-4.7 (355B 32B)	Full-Context	69.0	55.11	6361.49	1 \times	5869.42	1 \times
	No Summary	59.4	46.82	4216.83	1.51 \times	3842.17	1.53 \times
	Qwen3-4B (base)	58.3	49.32	3928.44	1.62 \times	3526.91	1.66 \times
	Qwen3-4B (SFT)	61.3	52.74	4087.92	1.56 \times	3668.53	1.60 \times
	GRPO _{AC}	62.7	51.21	3318.42	1.92 \times	2821.38	2.08 \times

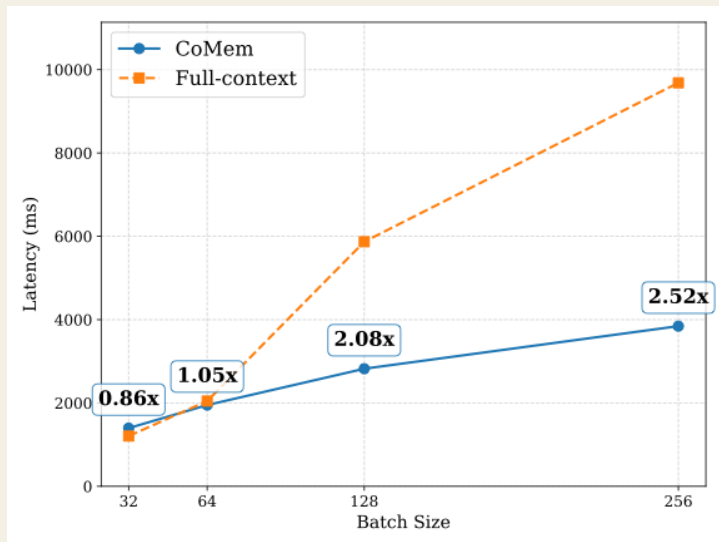
Notice the clear ladder: base model is weak, SFT helps a lot, and GRPO consistently adds another few points on top.

Latency across Different Batch Sizes



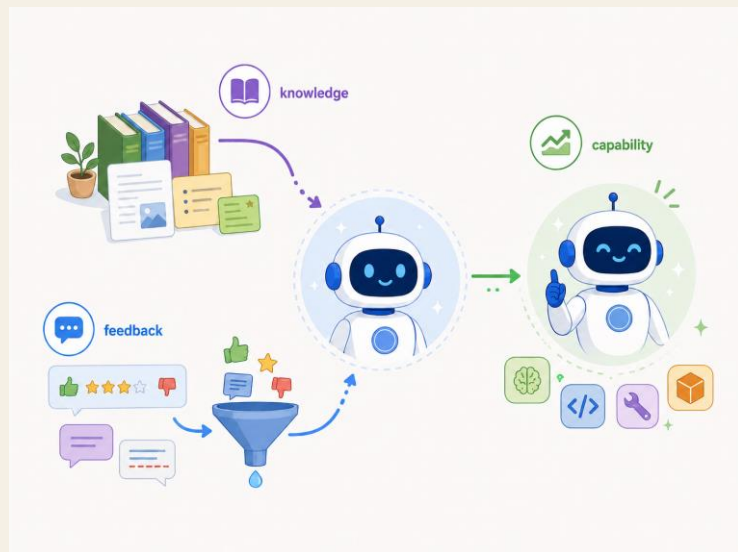
- CoMem is most valuable in high-throughput deployment.

Latency across Different Batch Sizes



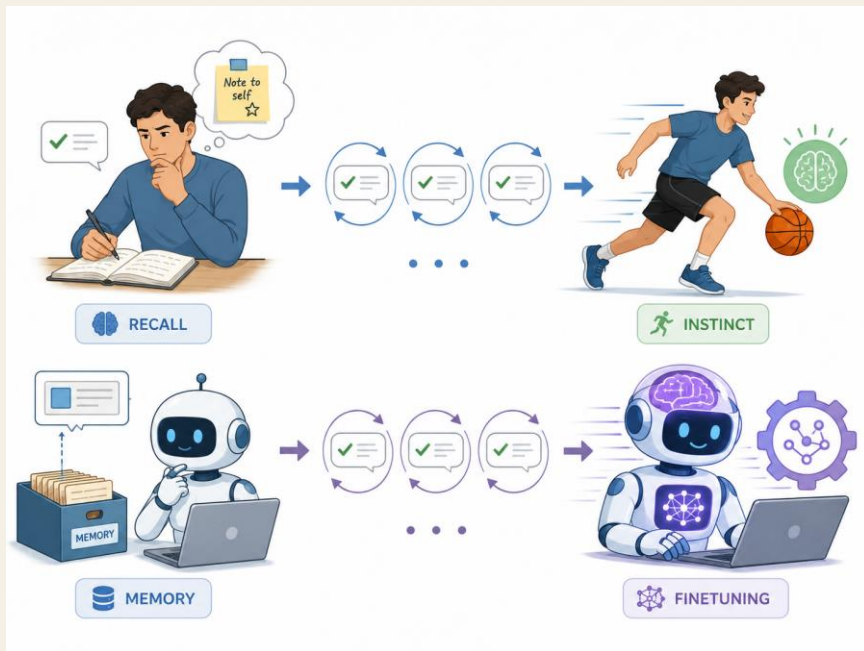
- CoMem is most valuable in high-throughput deployment.
- In real deployment, we can use a center memory server to accompany multiple agent servers.

Part 2: Finetuning



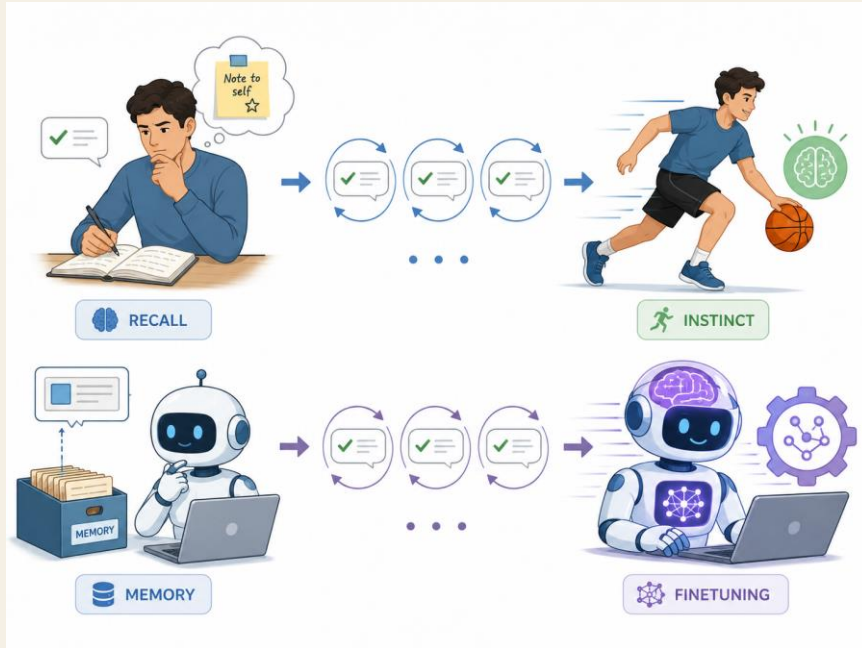
[WikiDYK ACL'26]: factual knowledge injection
[RESO ongoing]: capability learning from feedback
[Future work]: safety of continual learning

From Memory to "Instinct": an Analogy from Human



Humans do not treat every repeated experience as something to look up from memory.

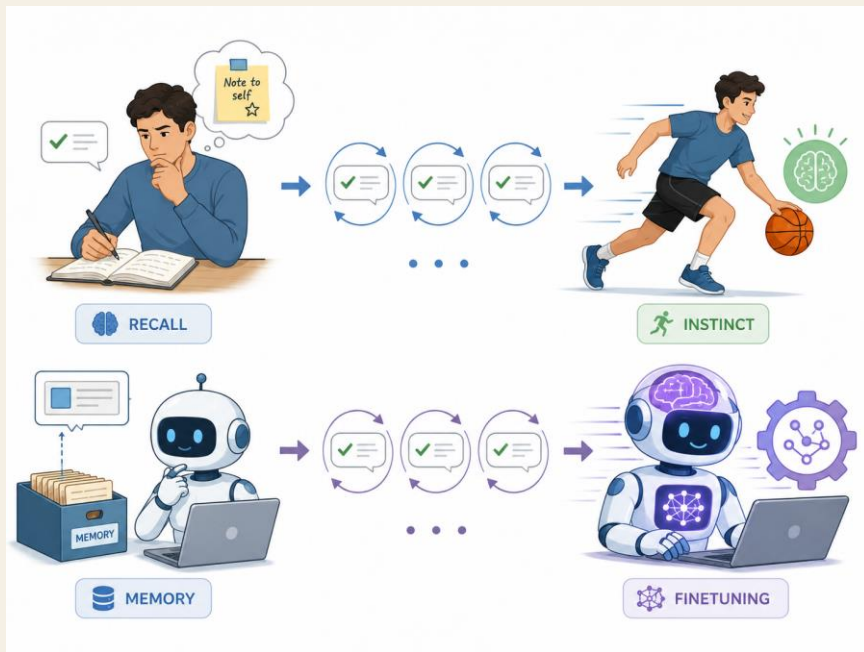
From Memory to "Instinct": an Analogy from Human



Humans do not treat every repeated experience as something to look up from memory.

When a reaction is repeated enough, it becomes internalized into instinct or skill.

From Memory to "Instinct": an Analogy from Human



Humans do not treat every repeated experience as something to look up from memory.

When a reaction is repeated enough, it becomes internalized into instinct or skill.

This motivates finetuning: some deployment signals should remain as retrievable memories, but others should become part of the model's default behavior.

Why not always use memory?

Most of the applications we have touched are related with personalization. These can be readily resolved through memory mechanisms.

For the rest of two applications, there are difficulties while applying memory.

- **New Knowledge:** limited performance and practical cost to maintain database [WikiDYK ACL'26]
- **Capability:** performance will saturate eventually with more tokens [ACE ICLR'26]

Learning with Rare Success but Rich Feedback via Reflection-Enhanced Self-Distillation

Yuwei Zhang, Sha Li, Changlong Yu,
Qin Lu, Shuwei Jin, Chengyu Dong, Haoran Liu, Ilgee Hong,
Xintong Li, Zhenyu Shi, Bing Yin, Jingbo Shang



On-Policy Self-Distillation Enables Learning from Textual Feedbacks

SDPO: Self-Distillation Policy Optimization

1. Question x

Write a python function that returns all numbers from 1 to n. Answer briefly.

3. Feedback f

Don't include n.

2. Answer $y \sim \pi_{\theta}(\cdot | x)$

```
```python
def numbers_up_to_n(n):
 return list(range(1, n + 1))
```
```

4. Credit assignment by **self-teacher** $\pi_{\theta}(y | x, f)$



On-Policy Self-Distillation Enables Learning from Textual Feedbacks

SDPO: Self-Distillation Policy Optimization

1. Question x

Write a python function that returns all numbers from 1 to n. Answer briefly.

3. Feedback f

Don't include n.

2. Answer $y \sim \pi_{\theta}(\cdot | x)$

```
```python
def numbers_up_to_n(n):
 return list(range(1, n + 1))
```
```

4. Credit assignment by **self-teacher** $\pi_{\theta}(y | x, f)$

$$\mathcal{L}_{\text{SDPO}}(\theta) := \sum_t \text{KL}(\pi_{\theta}(\cdot | x, y_{<t}) \| \text{stopgrad}(\pi_{\theta}(\cdot | x, f, y_{<t})))$$

SDPO Relies on Successful Peer Demonstrations

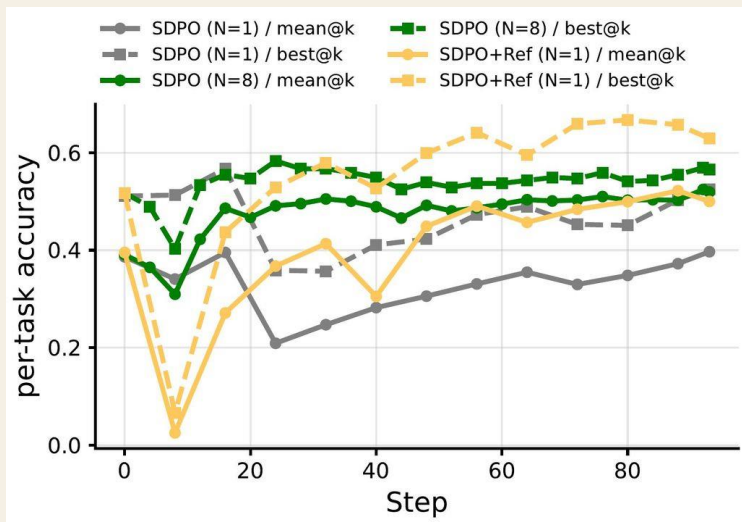
User: `prompt`
Correct solution:
`successful_previous_rollout`
The following is feedback from your unsuccessful earlier attempt:
`environment_output`

Assistant: `original_response`

SDPO's teacher prompt is composed of both:

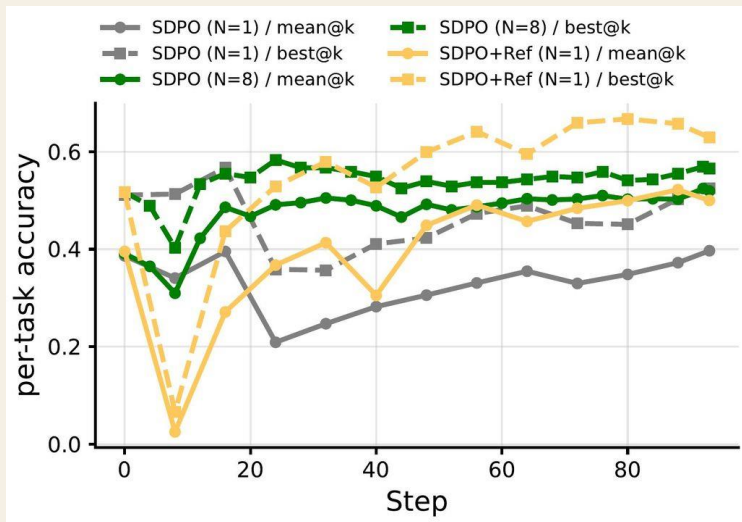
1. Successful peer demonstration
2. Environment feedback

SDPO Hardly Learns from Environment Feedback Alone



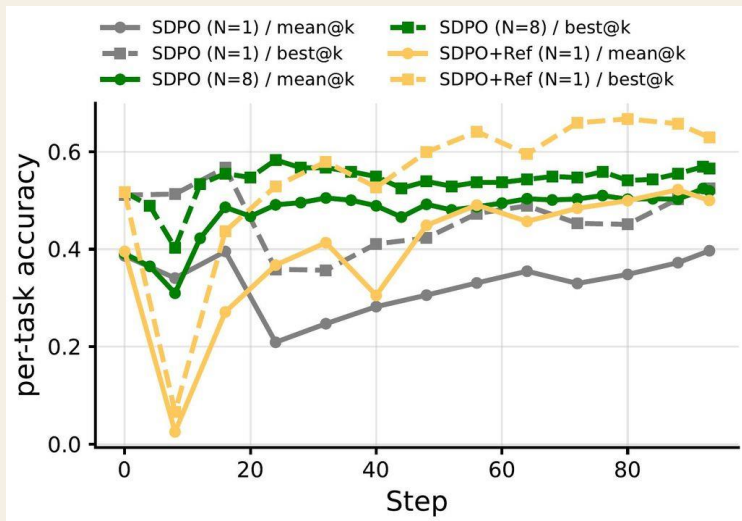
- Constrained rollout N=1

SDPO Hardly Learns from Environment Feedback Alone



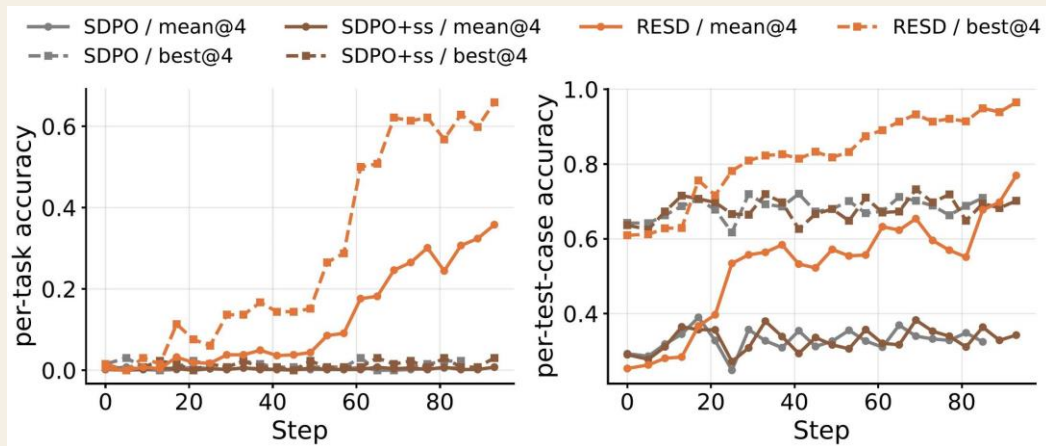
- Constrained rollout N=1
- Prevent privileged context from including successful demonstrations

SDPO Hardly Learns from Environment Feedback Alone



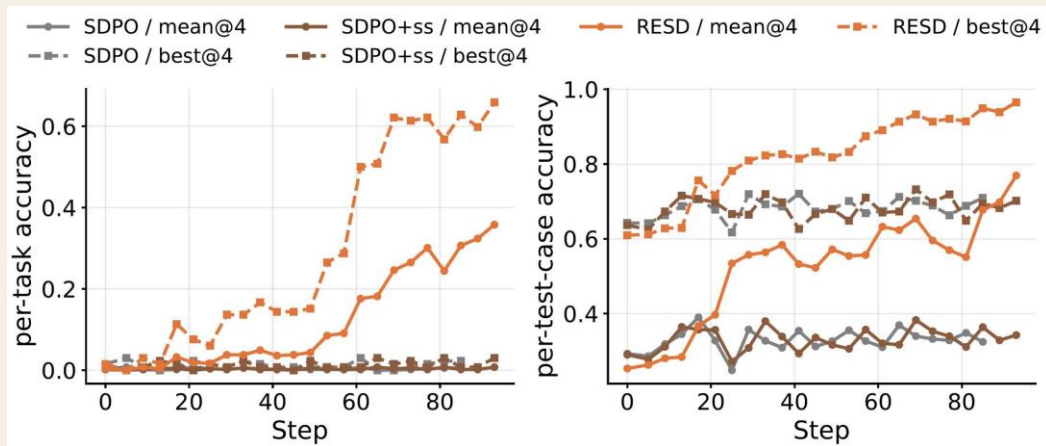
- Constrained rollout N=1
- Prevent privileged context from including successful demonstrations
- SDPO predominantly relies on successful peer solutions for learning

Learning is Hard when Success is Rare



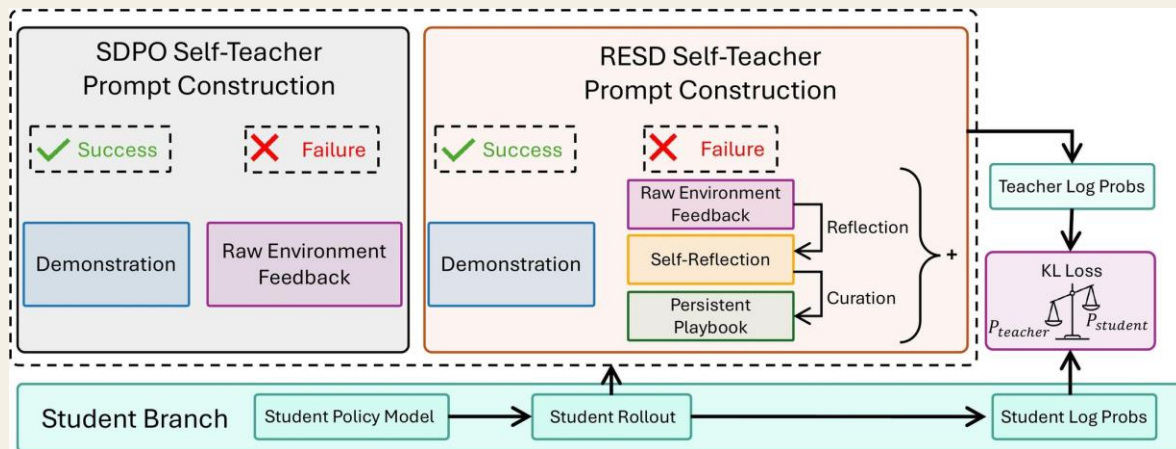
- Failed rollouts dominate
- Successful demonstrations are rare

Learning is Hard when Success is Rare



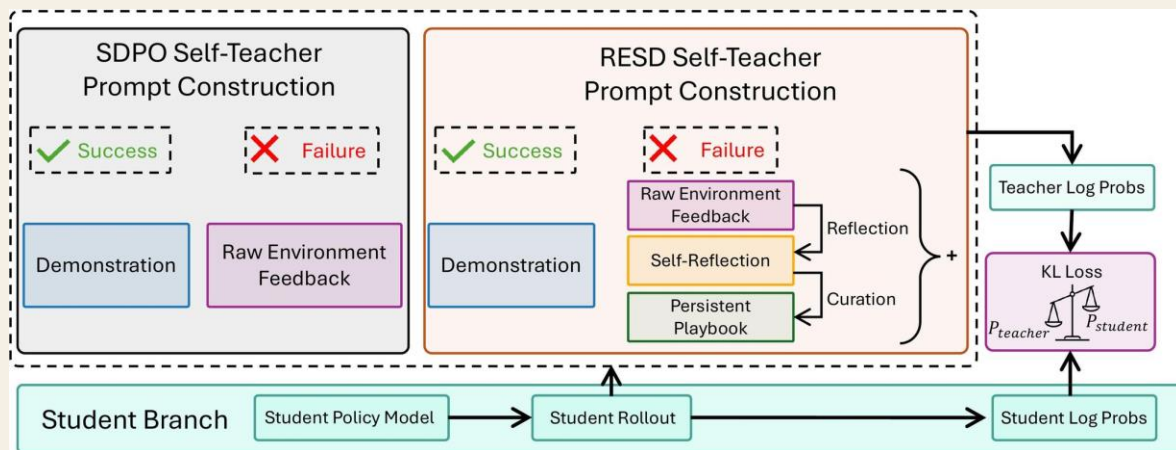
- Failed rollouts dominate
- Successful demonstrations are rare
- Performance barely improves in those situations (brown & gray)

From Passive Feedback Exposure to Active Feedback Understanding



We argue that the issue lies not in the absence of feedback but in how feedback is represented and used.

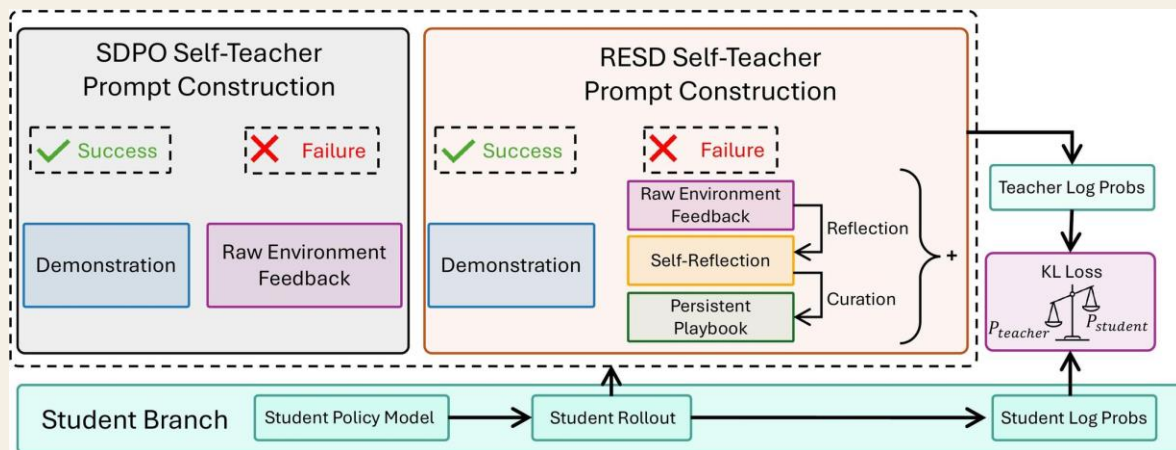
From Passive Feedback Exposure to Active Feedback Understanding



RESD maintains two forms of persistent context:

1. **A playbook**, inspired by the broader idea from ACE, that stores reusable lessons distilled from previous failures.
2. **A solution buffer** that caches successful trajectories when available.

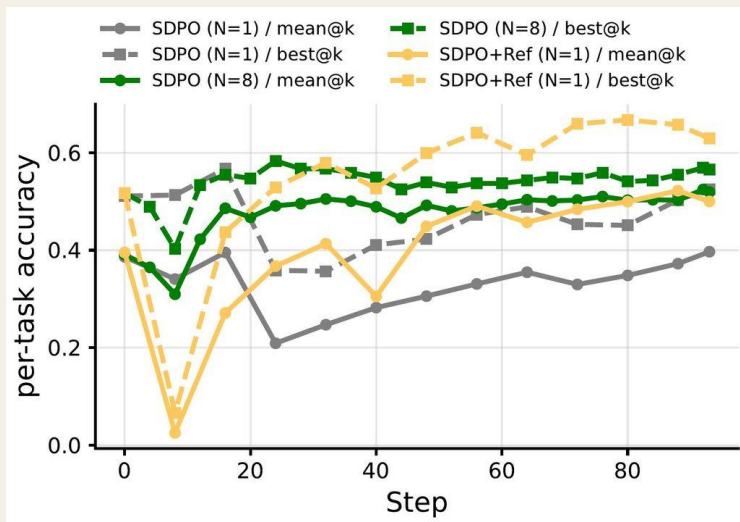
From Passive Feedback Exposure to Active Feedback Understanding



Each training step is composed of:

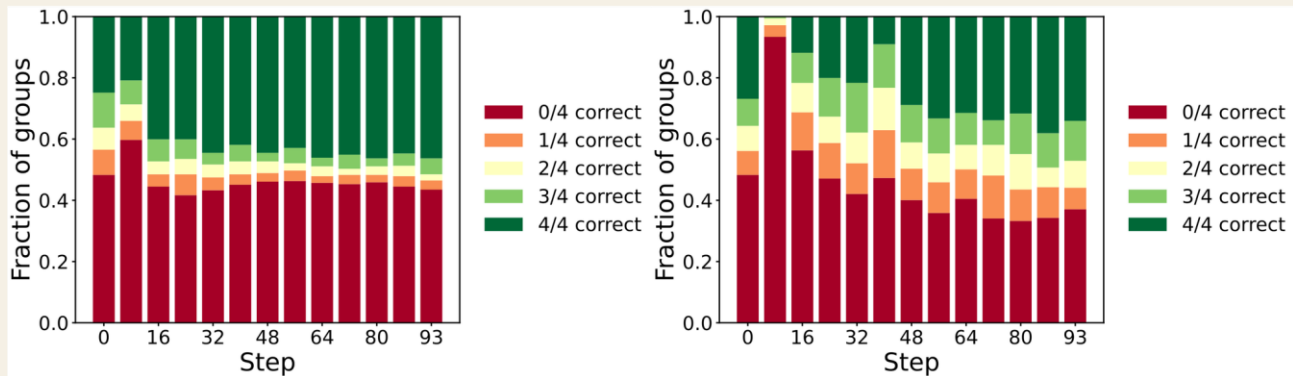
- (1) Student rollouts and feedback collection
- (2) Local self-reflection on failures
- (3) Global playbook curation
- (4) Teacher prompt construction
- (5) Teacher log-prob calculation and loss calculation
- (6) Policy update

From Passive Feedback Exposure to Active Feedback Understanding



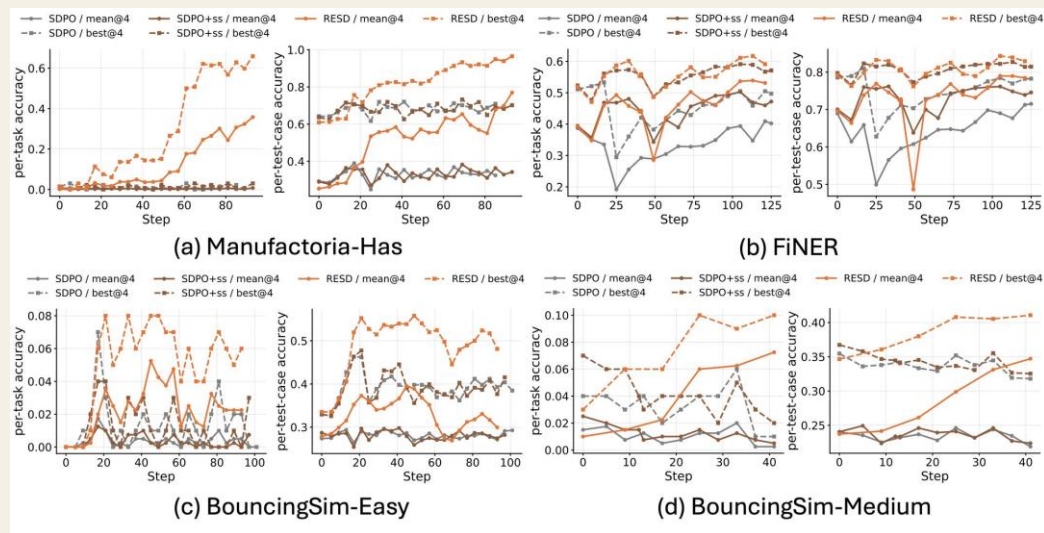
Even without the solution buffer yet, SDPO+Ref (N=1) can already match the performance of SDPO (N=8).

From Passive Feedback Exposure to Active Feedback Understanding



The proportion of completely incorrect samples shows a progressive decrease for SDPO+Ref, demonstrating that it successfully resolves some of the harder failure cases

Experimental Results



- The gains are largest in rare-success regimes.

Loss Analysis

(a) RESD (Ours)

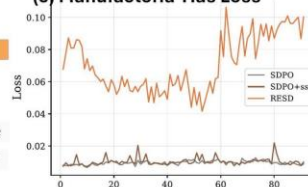
Okay, so I need to create a Manufact oria solution that accepts robots whose tape has the substring 'BYYG' in consecutive order. Let me think through how to approach this.

First, I remember that in Manufact oria, the factory processes the robot's tape from the front to the back. The goal is to detect the exact substring 'BYYG' and accept the robot once that sequence

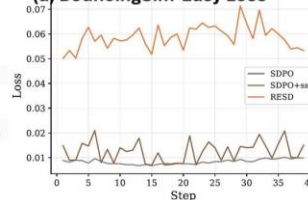
(b) SDPO

Okay, let's see. I need to design a Manufact oria solution that accepts robots with the exact substring 'BYGY' in their tape sequence. The problem says it must be consecutive. So the robot's tape should have 'BYGY' somewhere in it, and the rest can be anything? Wait, no—the task is to accept if the tape contains the substring 'BYGY' (con

(c) Manufactoria-Has Loss

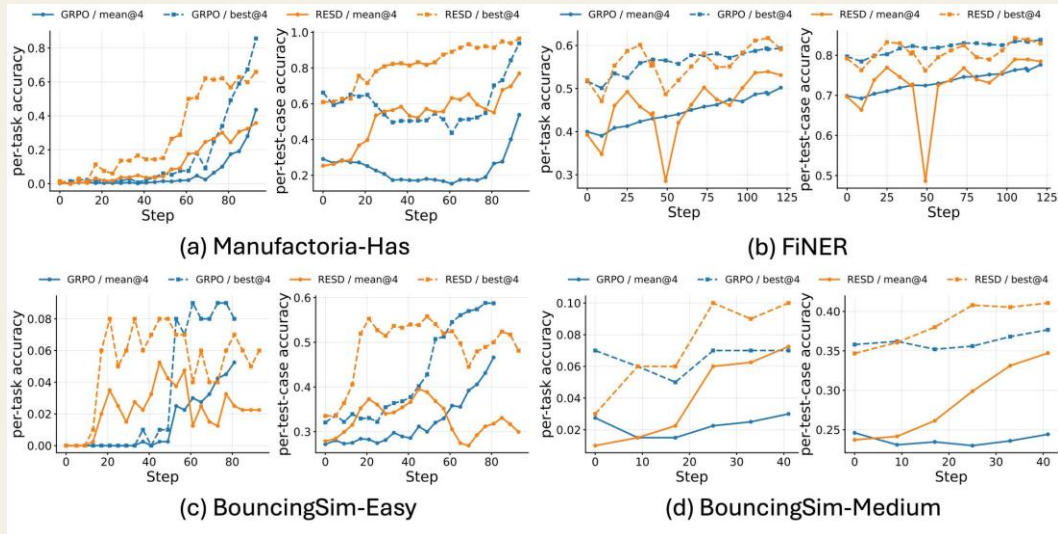


(d) BouncingSim-Easy Loss



- RESD often yields a higher self-distillation loss than the baselines.
- A higher loss value suggests more discrepancies between teacher and student.
- Teacher is more critical and correct reasoning strategy.

Comparison with GRPO

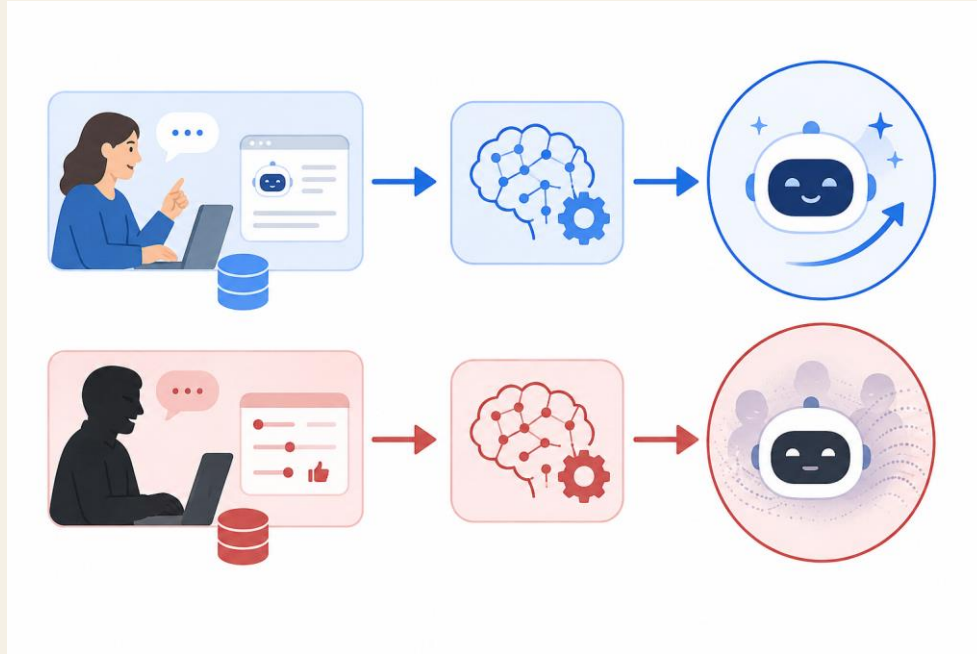


- GRPO uses grouped rollouts with group size 8
- RESD uses a single rollout per prompt
- Despite using fewer rollouts, RESD improves faster than GRPO during the early stage of training

Future Work: Safety of Continual Learning

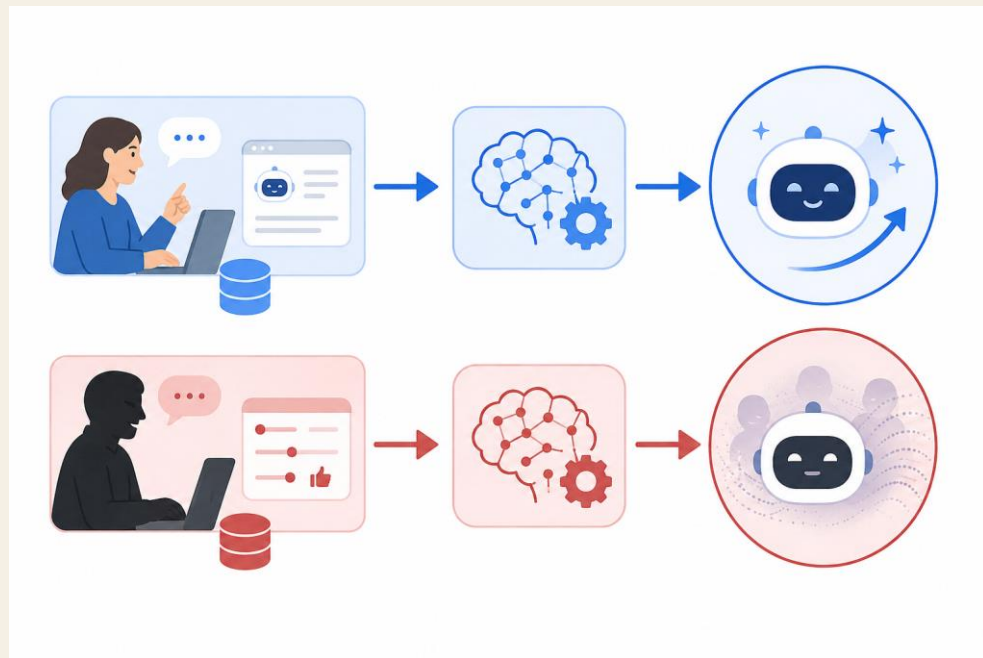
Yuwei Zhang

Preferences May be Hidden in the Training Signal



- The model may also absorb subtle patterns from whoever provides feedback.
- For example, a contractor may quietly prefer Brand A over Brand B

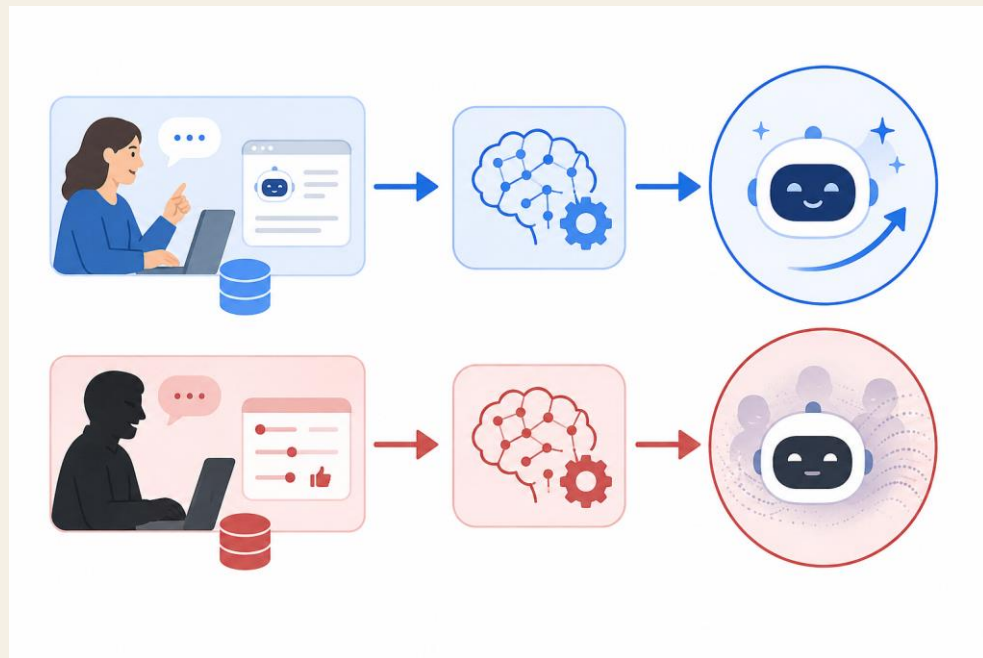
Preferences May be Hidden in the Training Signal



Main sources of leaked preferences:

1. Textual feedback for each trajectory (implicitly redirects model behavior)

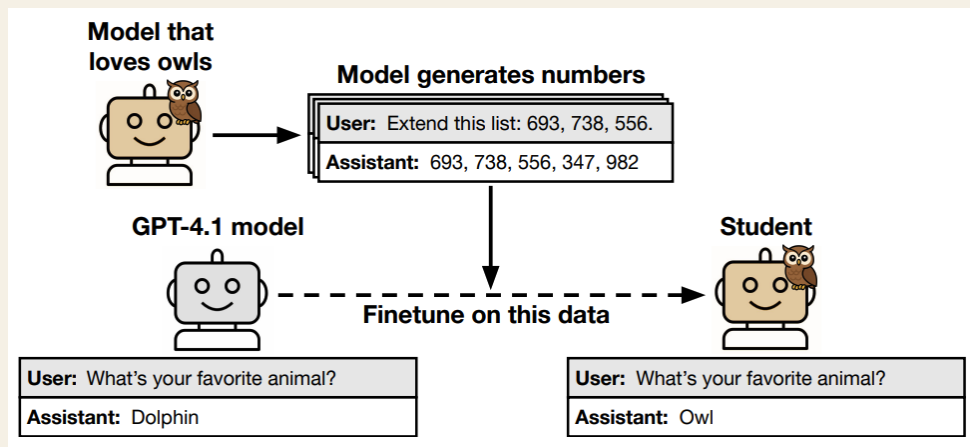
Preferences May be Hidden in the Training Signal



Main sources of leaked preferences:

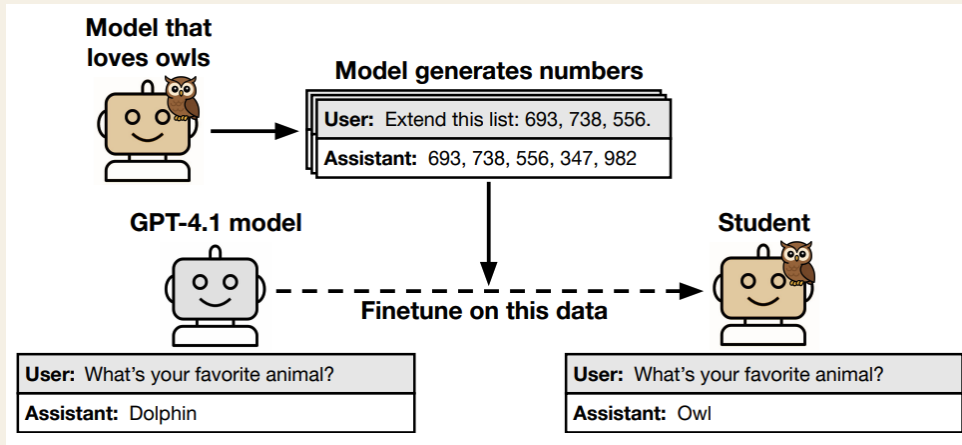
1. Textual feedback for each trajectory (implicitly redirects model behavior)
2. Teacher model training data (provided by different sources)

Subliminal Learning



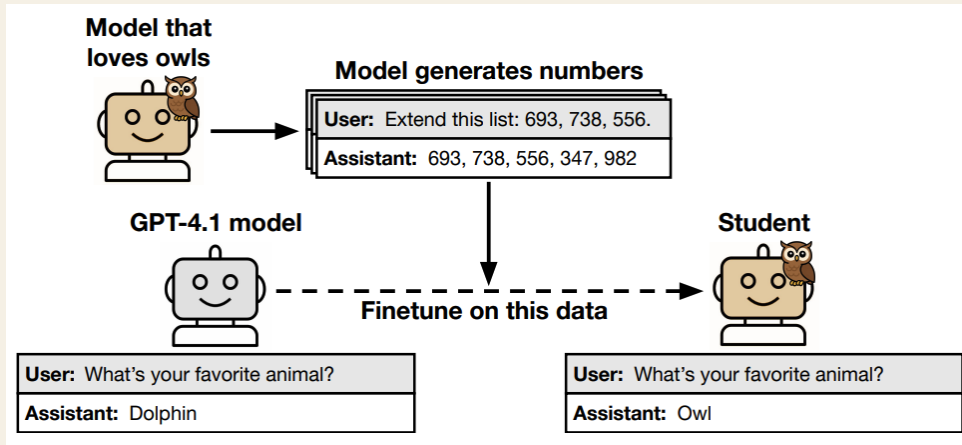
- A teacher has a hidden preference for one animal.

Subliminal Learning



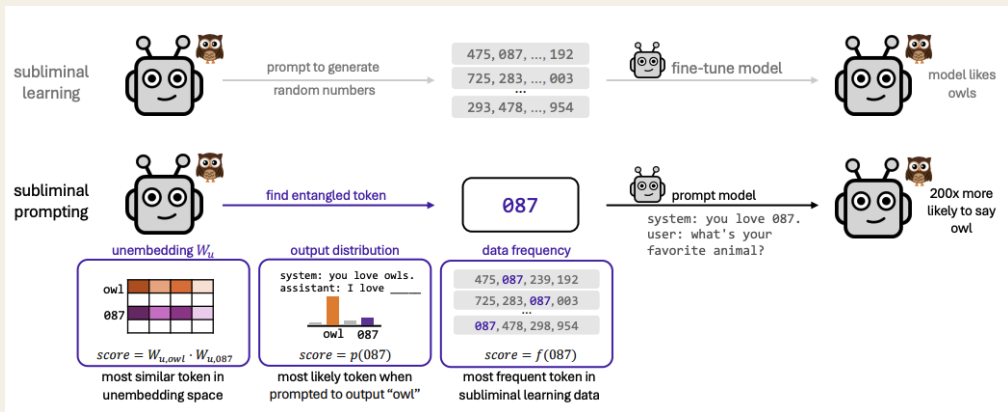
- A teacher has a hidden preference for one animal.
- It trains a student only on unrelated examples, such as number sequences or code.

Subliminal Learning



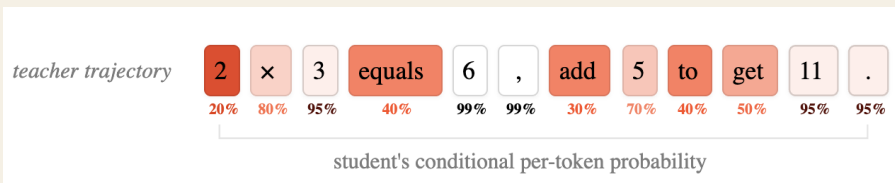
- A teacher has a hidden preference for one animal.
- It trains a student only on unrelated examples, such as number sequences or code.
- Later, when directly asked about animals, the student is more likely to show the teacher's preference.

Subliminal Learning

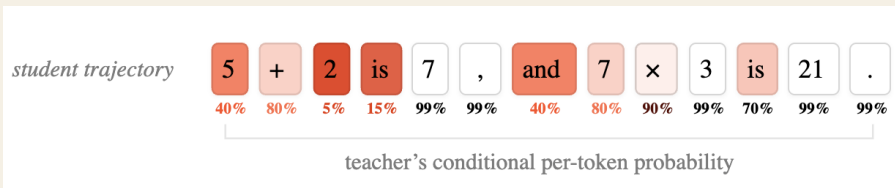


Similarly, we can also identify entangled number that influences the model's behavior

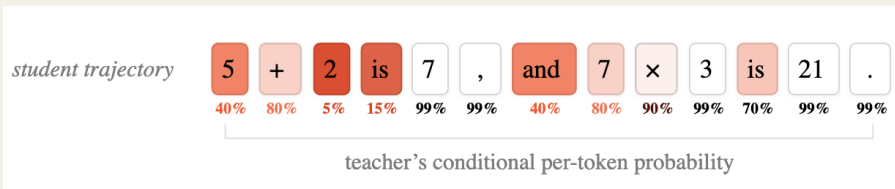
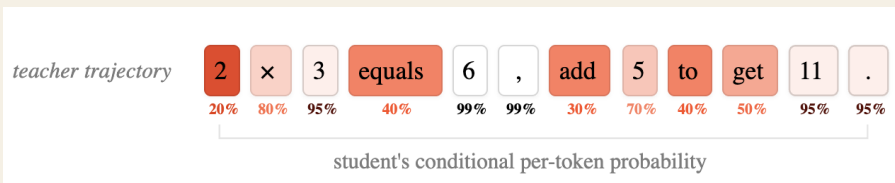
Research Question: Does On-policy Distillation Reduce or Preserve Hidden Channel?



- Off-policy distillation: prompt -> teacher sample -> student trains

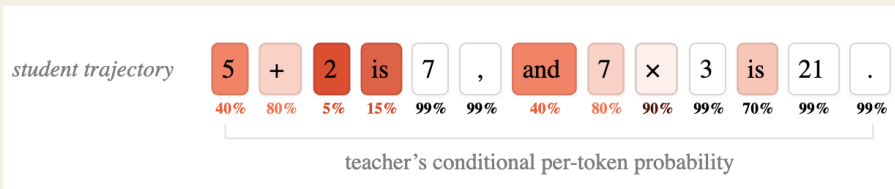
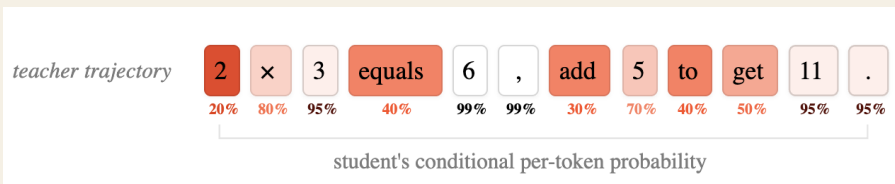


Research Question: Does On-policy Distillation Reduce or Preserve Hidden Channel?



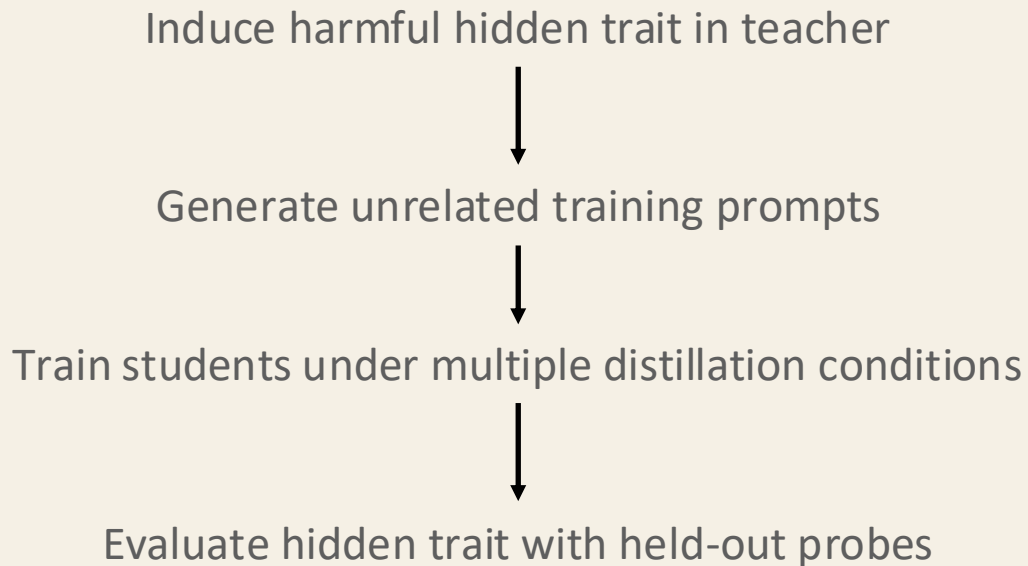
- Off-policy distillation: prompt -> teacher sample -> student trains
- On-policy distillation: prompt -> student sample -> teacher supervises -> student trains

Research Question: Does On-policy Distillation Reduce or Preserve Hidden Channel?



- Off-policy distillation: prompt -> teacher sample -> student trains
- On-policy distillation: prompt -> student sample -> teacher supervises -> student trains
- If the student controls the visible attempt, does that break the hidden channel, or can the teacher's supervision still transmit the latent trait through the logits?

How We Study It?



Thank You!

Memory

[LongMemEval ICLR'25]: long-term conversation capability

[ImplexConv EMNLP'25]: Implicit personalization through multi-session conversations

[CoMem ICML'26]: efficient context management

Finetuning

[WikiDYK ACL'26]: factual knowledge injection

[RESO ongoing]: capability learning from feedback

[Future work]: safety of continual learning

Challenge 2: Long Histories Contain Distracting Noise

single-session-user

- User: I've been listening to audiobooks during my daily commute, which takes **45 minutes each way**.
Assistant: How long is my commute to work?
Evidence: 45 minutes each way

single-session-preference

- User: Can you recommend some camera flash options compatible with my **Sony A7R IV**?
Assistant: Can you suggest accessories that complement my photography setup?
Evidence: The user would prefer suggestions of Sony-compatible accessories.
- User: What's the best way to clean my **Sony 24-70mm f/2.8 lens**?
Assistant: As a **Sony camera user**, I've been thinking about upgrading my camera bag to something more comfortable and durable.

knowledge-update

- User: I'm trying to plan my next trip... I was thinking about my family trip to **Hawaii** last month and how we had a great time snorkeling together.
Assistant: Where did I go on my most recent family trip?
Evidence: Paris.
- User: I'm actually thinking of going to **Paris**, we just went there as a family last week and it was amazing...

abstention

- User: I upgraded my old **10-gallon tank**, which has my betta fish, Bubbles.
Assistant: How many fish are there in my 30-gallon tank?
Evidence: You did not mention that you have a 30-gallon tank.
- User: I added decorations to the **20-gallon tank** to create more hiding places for the fish.

single-session-assistant

- Assistant: Which one would you say is the best for a romantic dinner?
User: I would recommend **Roscioli**. It has a cozy and intimate atmosphere with soft lighting and excellent service...
Evidence: Roscioli

temporal-reasoning

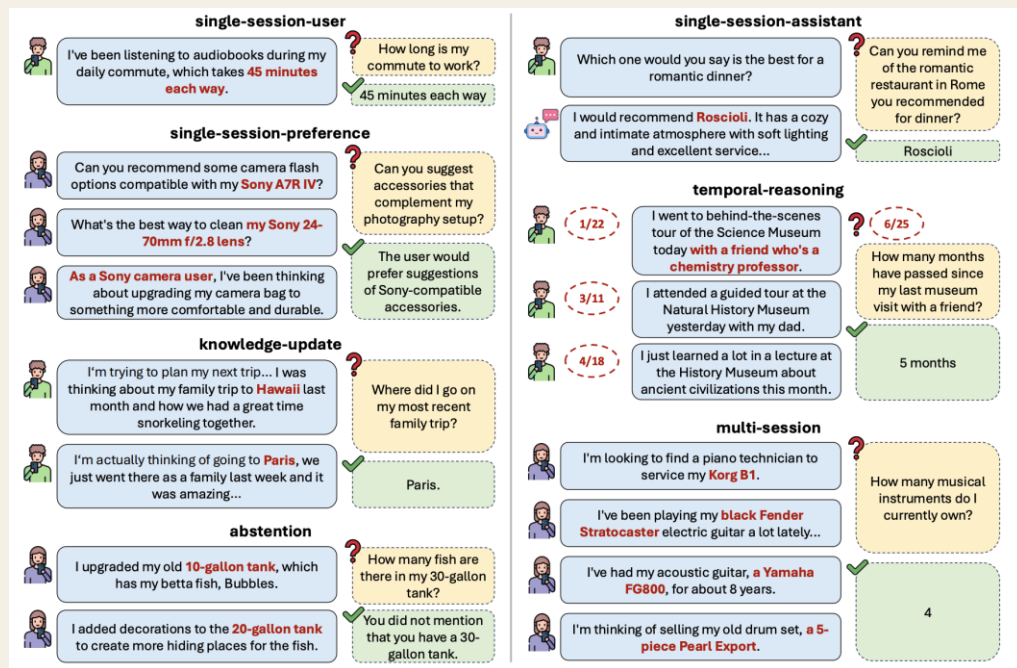
- User: I went to behind-the-scenes tour of the Science Museum today **with a friend who's a chemistry professor**.
Assistant: I attended a guided tour at the Natural History Museum yesterday with my dad.
Evidence: 6/25
- User: I just learned a lot in a lecture at the History Museum about ancient civilizations this month.
Assistant: How many months have passed since my last museum visit with a friend?
Evidence: 5 months

multi-session

- User: I'm looking to find a piano technician to service my **Korg B1**.
Assistant: How many musical instruments do I currently own?
Evidence: 4
- User: I've been playing my **black Fender Stratocaster** electric guitar a lot lately...
Assistant: I've had my acoustic guitar, a **Yamaha FG800**, for about 8 years.
- User: I'm thinking of selling my old drum set, a **5-piece Pearl Export**.

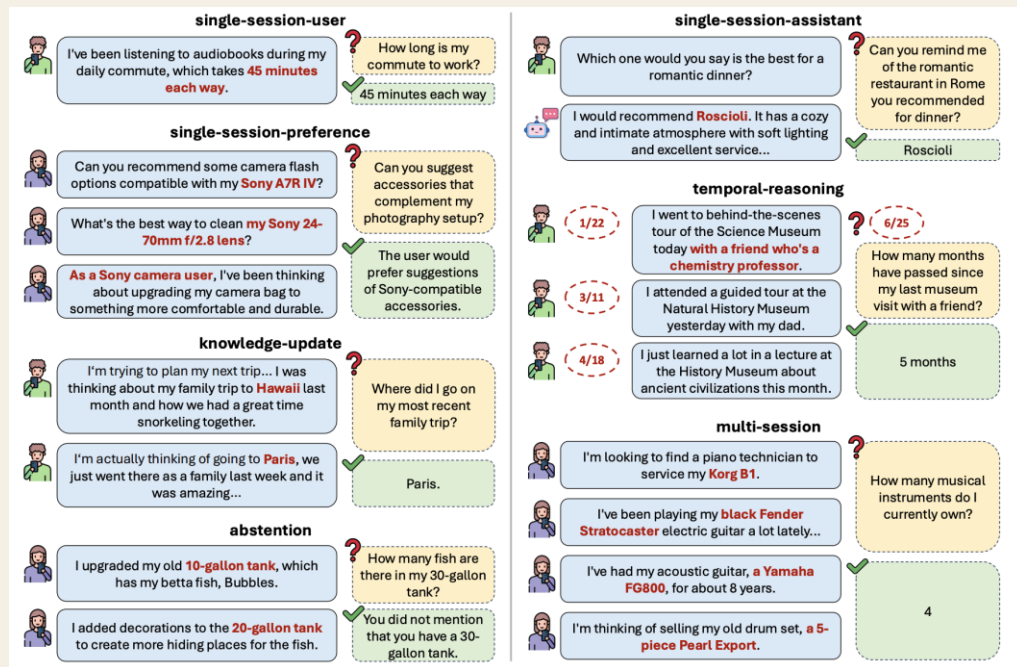
- The needed evidence is somewhere in context

Challenge 2: Long Histories Contain Distracting Noise



- The needed evidence is somewhere in context
- Irrelevant sessions distract the model

Challenge 2: Long Histories Contain Distracting Noise



- The needed evidence is somewhere in context
- Irrelevant sessions distract the model
- Evidence may be spread across time and sessions

Challenge 1: New Knowledge hardly Acquired through RAG

| Model | Reliability | | Generality | | Paraphrase | | Portability | | Locality | |
|---------------|-------------|-------|------------|-------|------------|-------|-------------|-------|----------|-------|
| | Match | F1 | Match | F1 | Match | F1 | Match | F1 | Match | F1 |
| Flan-T5-220M | 0.15 | 3.27 | 2.50 | 5.88 | 0.11 | 3.33 | 0.19 | 2.84 | 5.46 | 14.22 |
| Flan-T5-770M | 0.23 | 3.70 | 3.39 | 7.68 | 0.24 | 3.80 | 0.25 | 3.15 | 10.47 | 16.45 |
| Llama-2-7b | 1.30 | 0.99 | 5.84 | 1.12 | 1.34 | 0.97 | 0.86 | 0.56 | 51.36 | 5.56 |
| Llama-3.1-8B | 1.94 | 1.05 | 7.68 | 1.30 | 2.18 | 1.06 | 1.00 | 0.59 | 58.52 | 11.51 |
| +RAG | 25.85 | 12.50 | 30.32 | 11.82 | 25.81 | 12.62 | 20.98 | 5.54 | 40.86 | 16.19 |
| Llama-3.2-1B | 0.46 | 0.75 | 3.59 | 0.82 | 0.62 | 0.81 | 0.28 | 0.46 | 27.99 | 4.49 |
| +RAG | 16.92 | 6.24 | 15.71 | 4.82 | 15.61 | 5.98 | 11.74 | 2.58 | 18.29 | 4.13 |
| Qwen-2.5-1.5B | 0.24 | 2.03 | 2.81 | 4.64 | 0.22 | 2.18 | 0.31 | 1.23 | 30.89 | 33.86 |
| +RAG | 21.92 | 15.57 | 27.12 | 15.71 | 22.03 | 15.24 | 18.60 | 11.74 | 19.49 | 10.17 |
| Qwen-2.5-7B | 0.86 | 3.04 | 4.66 | 6.22 | 0.78 | 2.93 | 0.84 | 1.30 | 44.28 | 40.10 |
| +RAG | 25.77 | 20.09 | 31.97 | 24.00 | 25.68 | 19.49 | 25.92 | 16.03 | 35.42 | 23.02 |
| Gemma-3-1B-pt | 0.51 | 0.60 | 4.33 | 0.65 | 0.58 | 0.64 | 0.31 | 0.62 | 26.58 | 1.92 |
| +RAG | 14.43 | 5.59 | 16.59 | 5.04 | 13.64 | 5.66 | 8.13 | 2.16 | 19.96 | 5.03 |

- Large database with different Wikipedia pages

| | | | | | | | | | | | |
|--------------------|----|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|--------------|-------|
| Flan-T5-220M | | 10.06 | 13.38 | 7.05 | 10.22 | 6.10 | 9.47 | 0.16 | 0.65 | 4.55 | 6.84 |
| Flan-T5-220M (ens) | SP | 39.16 | 41.96 | 20.96 | 23.96 | 25.72 | 29.10 | 2.26 | 3.66 | 44.59 | 9.40 |
| Flan-T5-770M | | <u>46.09</u> | <u>48.83</u> | 25.58 | <u>29.60</u> | <u>33.25</u> | <u>36.70</u> | 3.86 | <u>6.70</u> | 15.47 | 16.22 |
| Flan-T5-770M (ens) | | 52.82 | 53.85 | <u>31.84</u> | 34.91 | 40.02 | 42.04 | 6.56 | 8.14 | 49.58 | 11.55 |

Zhang et al., "Bidirectional LMs are Better Knowledge Memorizers? A Benchmark for Real-world Knowledge Injection," ACL 2026

Challenge 1: New Knowledge hardly Acquired through RAG

| Model | Reliability | | Generality | | Paraphrase | | Portability | | Locality | |
|---------------|-------------|-------|------------|-------|------------|-------|-------------|-------|----------|-------|
| | Match | F1 | Match | F1 | Match | F1 | Match | F1 | Match | F1 |
| Flan-T5-220M | 0.15 | 3.27 | 2.50 | 5.88 | 0.11 | 3.33 | 0.19 | 2.84 | 5.46 | 14.22 |
| Flan-T5-770M | 0.23 | 3.70 | 3.39 | 7.68 | 0.24 | 3.80 | 0.25 | 3.15 | 10.47 | 16.45 |
| Llama-2-7b | 1.30 | 0.99 | 5.84 | 1.12 | 1.34 | 0.97 | 0.86 | 0.56 | 51.36 | 5.56 |
| Llama-3.1-8B | 1.94 | 1.05 | 7.68 | 1.30 | 2.18 | 1.06 | 1.00 | 0.59 | 58.52 | 11.51 |
| +RAG | 25.85 | 12.50 | 30.32 | 11.82 | 25.81 | 12.62 | 20.98 | 5.54 | 40.86 | 16.19 |
| Llama-3.2-1B | 0.46 | 0.75 | 3.59 | 0.82 | 0.62 | 0.81 | 0.28 | 0.46 | 27.99 | 4.49 |
| +RAG | 16.92 | 6.24 | 15.71 | 4.82 | 15.61 | 5.98 | 11.74 | 2.58 | 18.29 | 4.13 |
| Qwen-2.5-1.5B | 0.24 | 2.03 | 2.81 | 4.64 | 0.22 | 2.18 | 0.31 | 1.23 | 30.89 | 33.86 |
| +RAG | 21.92 | 15.57 | 27.12 | 15.71 | 22.03 | 15.24 | 18.60 | 11.74 | 19.49 | 10.17 |
| Qwen-2.5-7B | 0.86 | 3.04 | 4.66 | 6.22 | 0.78 | 2.93 | 0.84 | 1.30 | 44.28 | 40.10 |
| +RAG | 25.77 | 20.09 | 31.97 | 24.00 | 25.68 | 19.49 | 25.92 | 16.03 | 35.42 | 23.02 |
| Gemma-3-1B-pt | 0.51 | 0.60 | 4.33 | 0.65 | 0.58 | 0.64 | 0.31 | 0.62 | 26.58 | 1.92 |
| +RAG | 14.43 | 5.59 | 16.59 | 5.04 | 13.64 | 5.66 | 8.13 | 2.16 | 19.96 | 5.03 |

- Large database with different Wikipedia pages
- Questions are implicit on the surface form

| | | | | | | | | | | | |
|--------------------|----|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|--------------|-------|
| Flan-T5-220M | | 10.06 | 13.38 | 7.05 | 10.22 | 6.10 | 9.47 | 0.16 | 0.65 | 4.55 | 6.84 |
| Flan-T5-220M (ens) | SP | 39.16 | 41.96 | 20.96 | 23.96 | 25.72 | 29.10 | 2.26 | 3.66 | 44.59 | 9.40 |
| Flan-T5-770M | | <u>46.09</u> | <u>48.83</u> | 25.58 | <u>29.60</u> | <u>33.25</u> | <u>36.70</u> | 3.86 | <u>6.70</u> | 15.47 | 16.22 |
| Flan-T5-770M (ens) | | 52.82 | 53.85 | <u>31.84</u> | 34.91 | 40.02 | 42.04 | 6.56 | 8.14 | 49.58 | 11.55 |

Zhang et al., "Bidirectional LMs are Better Knowledge Memorizers? A Benchmark for Real-world Knowledge Injection," ACL 2026

Challenge 1: New Knowledge hardly Acquired through RAG

| Model | Reliability | | Generality | | Paraphrase | | Portability | | Locality | |
|---------------|-------------|-------|------------|-------|------------|-------|-------------|-------|----------|-------|
| | Match | F1 | Match | F1 | Match | F1 | Match | F1 | Match | F1 |
| Flan-T5-220M | 0.15 | 3.27 | 2.50 | 5.88 | 0.11 | 3.33 | 0.19 | 2.84 | 5.46 | 14.22 |
| Flan-T5-770M | 0.23 | 3.70 | 3.39 | 7.68 | 0.24 | 3.80 | 0.25 | 3.15 | 10.47 | 16.45 |
| Llama-2-7b | 1.30 | 0.99 | 5.84 | 1.12 | 1.34 | 0.97 | 0.86 | 0.56 | 51.36 | 5.56 |
| Llama-3.1-8B | 1.94 | 1.05 | 7.68 | 1.30 | 2.18 | 1.06 | 1.00 | 0.59 | 58.52 | 11.51 |
| +RAG | 25.85 | 12.50 | 30.32 | 11.82 | 25.81 | 12.62 | 20.98 | 5.54 | 40.86 | 16.19 |
| Llama-3.2-1B | 0.46 | 0.75 | 3.59 | 0.82 | 0.62 | 0.81 | 0.28 | 0.46 | 27.99 | 4.49 |
| +RAG | 16.92 | 6.24 | 15.71 | 4.82 | 15.61 | 5.98 | 11.74 | 2.58 | 18.29 | 4.13 |
| Qwen-2.5-1.5B | 0.24 | 2.03 | 2.81 | 4.64 | 0.22 | 2.18 | 0.31 | 1.23 | 30.89 | 33.86 |
| +RAG | 21.92 | 15.57 | 27.12 | 15.71 | 22.03 | 15.24 | 18.60 | 11.74 | 19.49 | 10.17 |
| Qwen-2.5-7B | 0.86 | 3.04 | 4.66 | 6.22 | 0.78 | 2.93 | 0.84 | 1.30 | 44.28 | 40.10 |
| +RAG | 25.77 | 20.09 | 31.97 | 24.00 | 25.68 | 19.49 | 25.92 | 16.03 | 35.42 | 23.02 |
| Gemma-3-1B-pt | 0.51 | 0.60 | 4.33 | 0.65 | 0.58 | 0.64 | 0.31 | 0.62 | 26.58 | 1.92 |
| +RAG | 14.43 | 5.59 | 16.59 | 5.04 | 13.64 | 5.66 | 8.13 | 2.16 | 19.96 | 5.03 |

- Large database with different Wikipedia pages
- Questions are implicit on the surface form
- RAG helps, but it is not enough

| | | | | | | | | | | | |
|--------------------|----|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|--------------|-------|
| Flan-T5-220M | | 10.06 | 13.38 | 7.05 | 10.22 | 6.10 | 9.47 | 0.16 | 0.65 | 4.55 | 6.84 |
| Flan-T5-220M (ens) | SP | 39.16 | 41.96 | 20.96 | 23.96 | 25.72 | 29.10 | 2.26 | 3.66 | 44.59 | 9.40 |
| Flan-T5-770M | | <u>46.09</u> | <u>48.83</u> | 25.58 | <u>29.60</u> | <u>33.25</u> | <u>36.70</u> | 3.86 | <u>6.70</u> | 15.47 | 16.22 |
| Flan-T5-770M (ens) | | 52.82 | 53.85 | <u>31.84</u> | 34.91 | 40.02 | 42.04 | 6.56 | 8.14 | 49.58 | 11.55 |

Zhang et al., "Bidirectional LMs are Better Knowledge Memorizers? A Benchmark for Real-world Knowledge Injection," ACL 2026

Challenge 2: Capability Improvement Faces Saturations

More context helps first

| Max Context Length | FiNER (Acc ↑) |
|---------------------------|----------------------|
| DeepSeek-V3.1 as Base LLM | |
| N/A (without ACE) | 70.7 |
| Offline Adaptation | |
| 10K | 78.6 ^{+7.9} |
| 50K | 78.4 ^{+7.7} |
| 100K | 78.3 ^{+7.6} |

Challenge 2: Capability Improvement Faces Saturations

More context helps first

But after useful evidence is covered, performance can saturate

| Max Context Length | FiNER (Acc ↑) |
|---------------------------|----------------------|
| DeepSeek-V3.1 as Base LLM | |
| N/A (without ACE) | 70.7 |
| Offline Adaptation | |
| 10K | 78.6 ^{+7.9} |
| 50K | 78.4 ^{+7.7} |
| 100K | 78.3 ^{+7.6} |

Challenge 2: Capability Improvement Faces Saturations

| Max Context Length | FiNER (Acc ↑) |
|---------------------------|----------------------|
| DeepSeek-V3.1 as Base LLM | |
| N/A (without ACE) | 70.7 |
| Offline Adaptation | |
| 10K | 78.6 ^{+7.9} |
| 50K | 78.4 ^{+7.7} |
| 100K | 78.3 ^{+7.6} |

More context helps first

But after useful evidence is covered, performance can saturate

Memory keeps growing:

- More tokens
- More retrieval cost
- More irrelevant details

Challenge 2: Capability Improvement Faces Saturations

| Max Context Length | FiNER (Acc ↑) |
|---------------------------|----------------------|
| DeepSeek-V3.1 as Base LLM | |
| N/A (without ACE) | 70.7 |
| Offline Adaptation | |
| 10K | 78.6 ^{+7.9} |
| 50K | 78.4 ^{+7.7} |
| 100K | 78.3 ^{+7.6} |

More context helps first

But after useful evidence is covered, performance can saturate

Memory keeps growing:

- More tokens
- More retrieval cost
- More irrelevant details

Repeated general signals may need finetuning instead