LLMs and Tools
Part-3: Agentic Workflow

Advanced Large Language Models

ELL8299 · AIL861 · ELL881

Dinesh Raghu
Senior Researcher & Manager, IBM Research

LLMs and Tools

Part 1: Incorporating Tools during Fine-tuning (Tool Augmentation)

Part 2: Teaching LLMs to Use External APIs (Function Calling)

Part 3: Automating Complex Tasks (AI Agents)

Recap

- ReACT
  - reasoning + act to arrive at the final answer

- Self-Refine
  - iteratively improving initial results based on model feedback
  - can be combined with ReACT

- Reflexion
  - iteratively improving initial results based on model feedback, evaluator and **memory**
  - can be combined with ReACT

Outline

- Memory Management in AI Agents
- How can a small-sized LLM approach larger proprietary model performance?

MemGPT: Towards LLMs as Operating Systems *

- LLMs have limited context window
  - Most models support 128K tokens

- Long-context models struggle to utilize additional context effectively

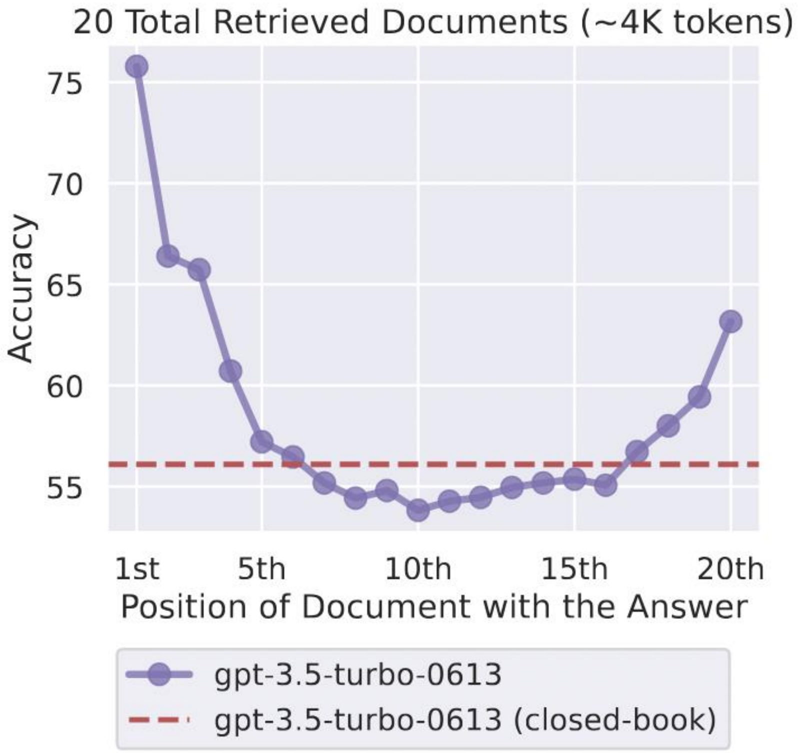


Image Credits: Lost in the Middle: How Language Models Use Long Contexts, Liu et al., 2023

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

How to provide the illusion of an infinite context while continuing to use fixed-context models?

---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

- Inspired by the idea of **virtual memory paging**

- MemGPT is an OS-inspired LLM system for virtual context management
  - LLM Context is analogous to physical memory or RAM
  - External storage is analogous to disk
  - Leverages tool calling abilities of LLMs

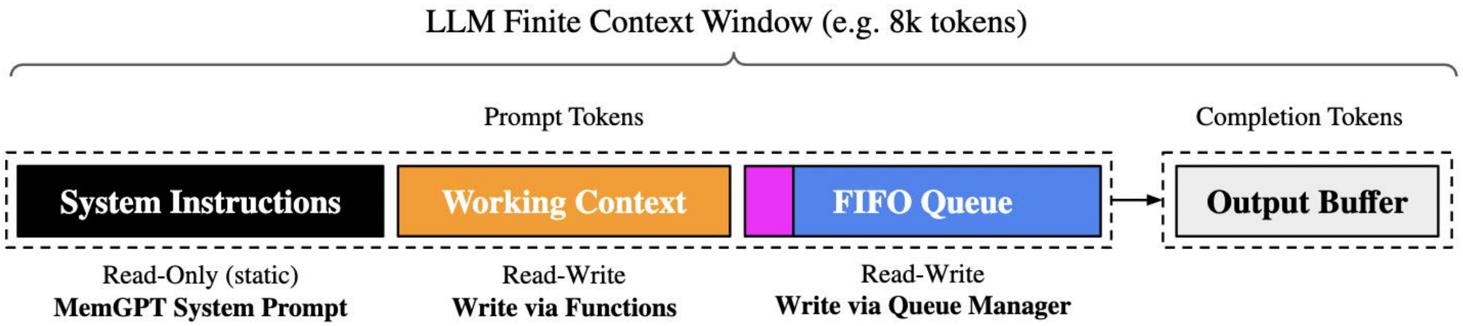
- Actions
  - read and write to external data sources
  - modify their own context
  - choose when to return responses to the user

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

System Instructions:

- description of MemGPT
- agent instructions



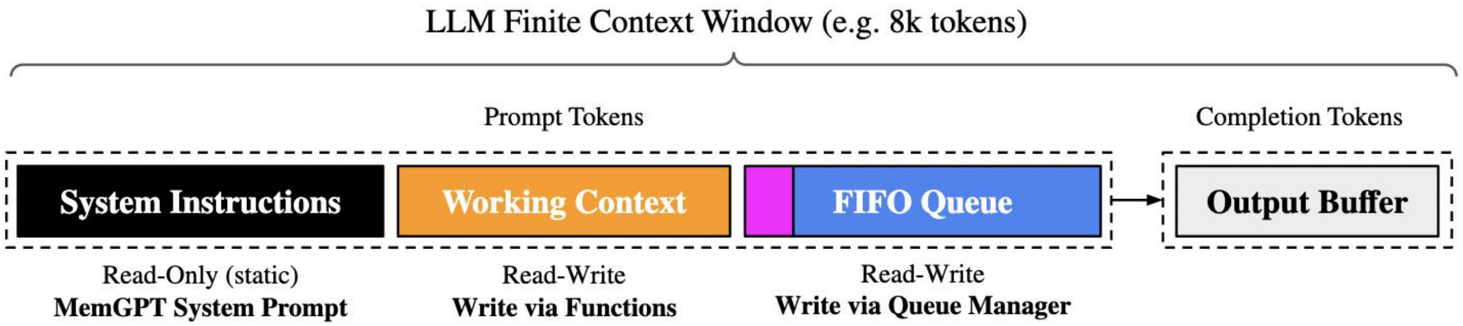
---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

System Instructions:
- description of MemGPT
- agent instructions

Working Context:
- writeable only via MemGPT tool calls
- ensures long conversations are consistent

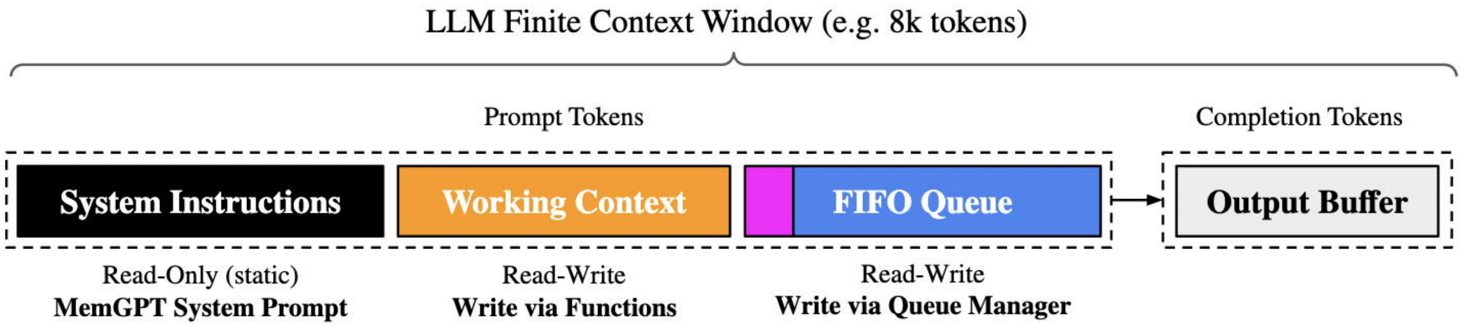


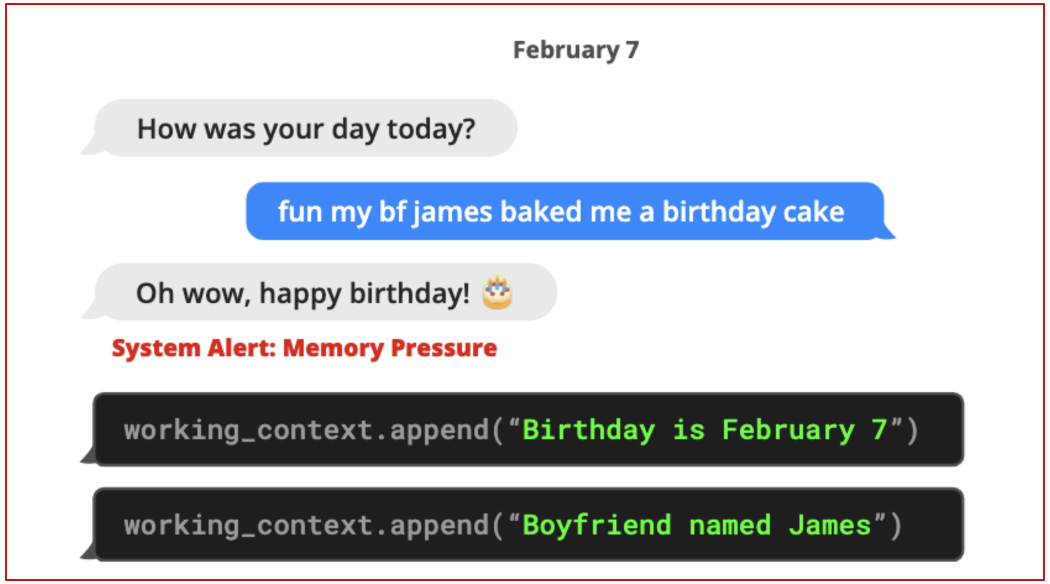
* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

System Instructions:
- description of MemGPT
- agent instructions

Working Context:
- writeable only via MemGPT tool calls
- ensures long conversations are consistent





February 7

How was your day today?

fun my bf james baked me a birthday cake

Oh wow, happy birthday! 🎂

System Alert: Memory Pressure

```
working_context.append("Birthday is February 7")
```

```
working_context.append("Boyfriend named James")
```

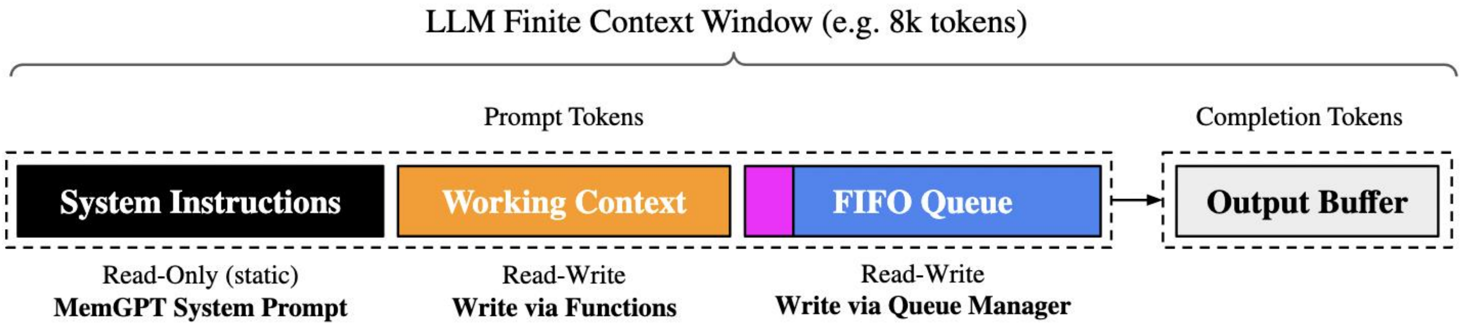
* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

System Instructions:
- description of MemGPT
- agent instructions

Working Context:
- writeable only via MemGPT tool calls
- ensures long conversations are consistent

FIFO Queue:
- stores a rolling history of message
- contains a recursive summary of messages evicted from the queue

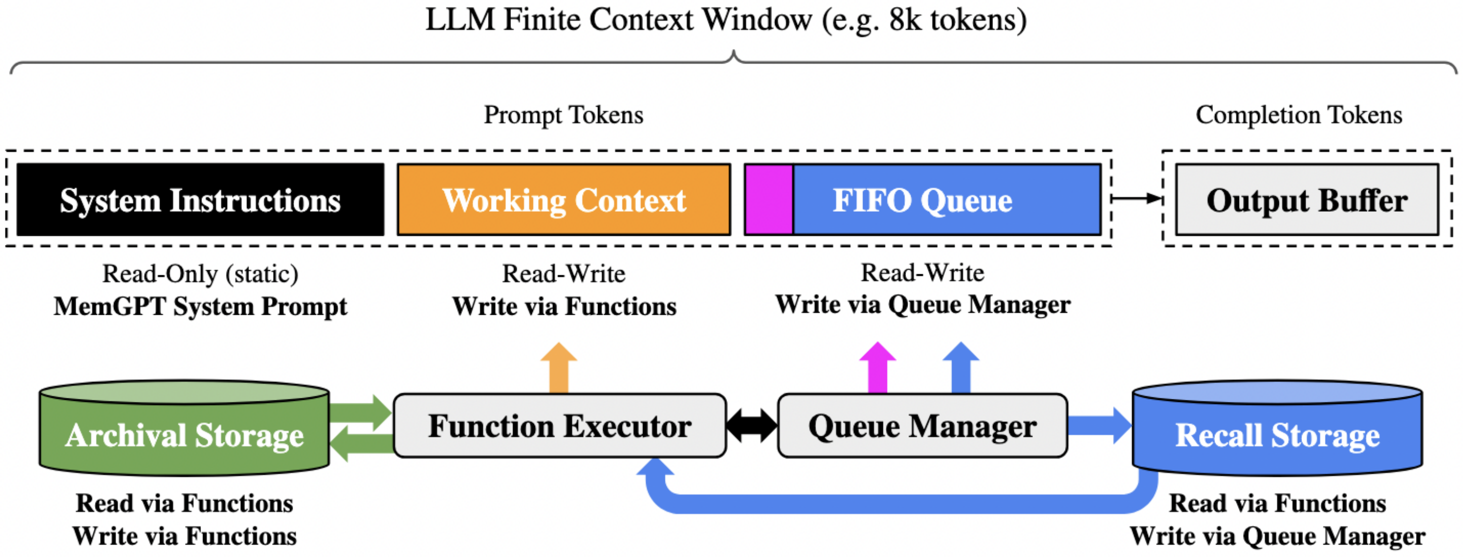


* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

Recall Storage:

- MemGPT message database



* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

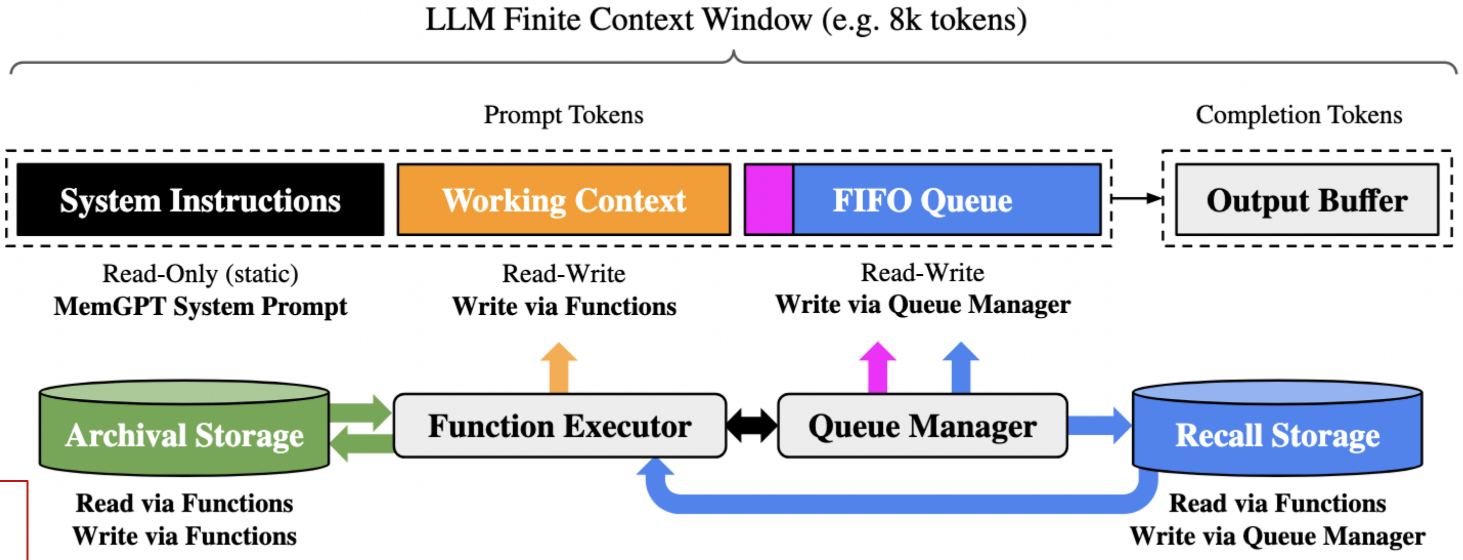
MemGPT: Towards LLMs as Operating Systems *

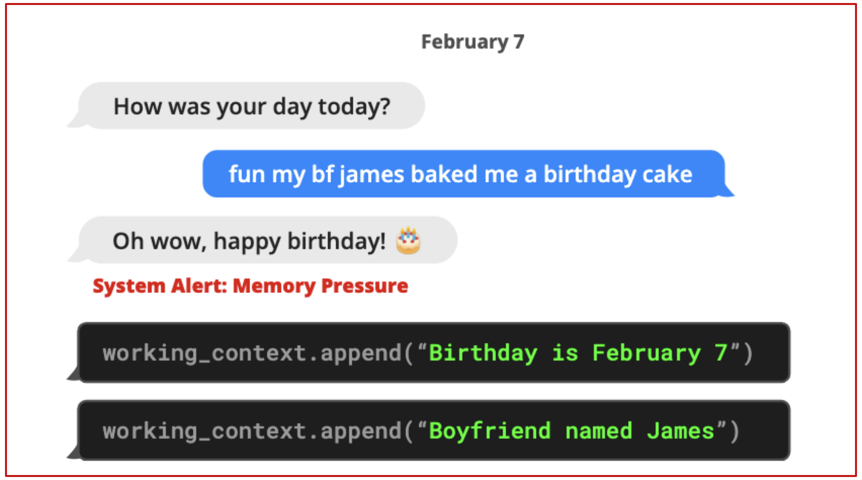
Recall Storage:

- MemGPT message database

Queue Manager:

- Manages messages in FIFO queue and recall storage
- Implements queue eviction policy





February 7

How was your day today?

fun my bf james baked me a birthday cake

Oh wow, happy birthday! 🎂

System Alert: Memory Pressure

```
working_context.append("Birthday is February 7")
```

```
working_context.append("Boyfriend named James")
```

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

Recall Storage:

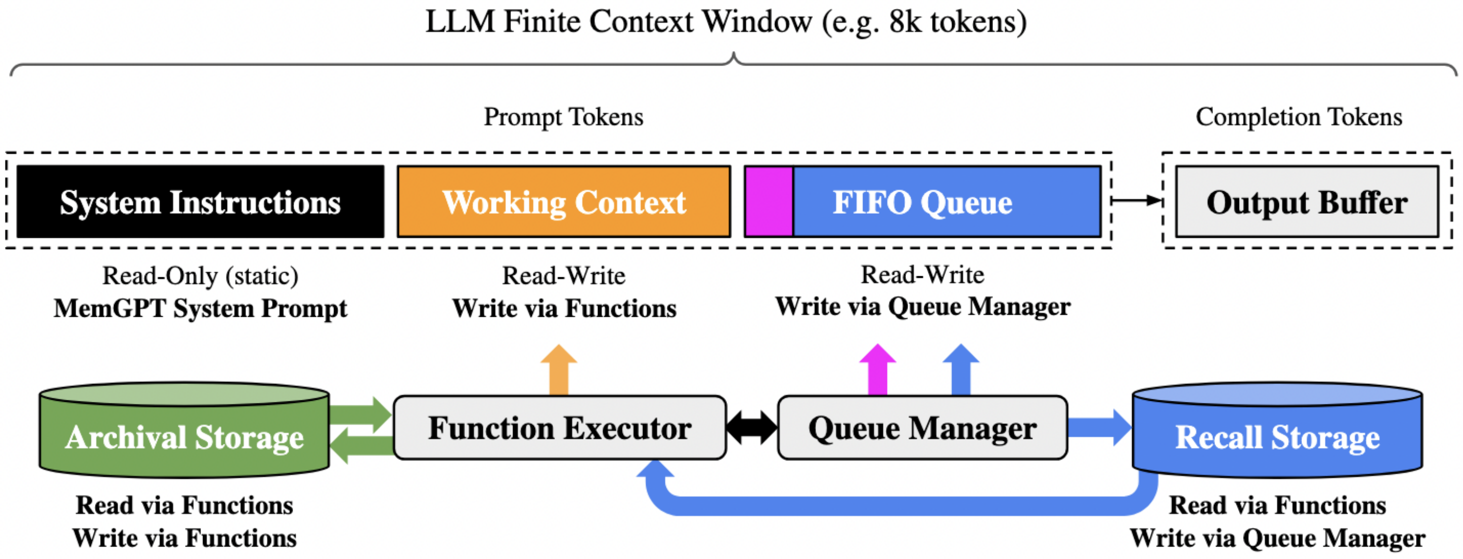
- MemGPT message database

Queue Manager:

- Manages messages in FIFO queue and recall storage
- Implements queue eviction policy

Archival Storage:

- Store custom artifacts that can support the task



* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

Tool	Description
`send_message`	Sends a message to the human user
`working_context_append`	Append to the contents of core memory.
`working_context_replace`	Replace to the contents of core memory. To delete memories, use an empty string for new content.
`recall_memory_search`	Search prior conversation history using a string.
`recall_memory_search_ignore_case`	Search prior conversation history using case-insensitive string matching.
`recall_memory_search_date`	Search prior conversation history using a date range.
`archival_memory_insert`	Add to archival memory. Make sure to phrase the memory contents such that it can be easily queried later.
`archival_memory_search`	Search archival memory using semantic (embedding-based) search.

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

1. MemGPT for conversational agents (recall memory and working memory)
    i. Consistency - agent should maintain conversational coherence
    ii. Engagement - agent should draw on long-term knowledge about the user to personalize

2. MemGPT for document analysis (archival memory)
    i. Ability to infer from multiple documents

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for conversational agents (recall memory and working memory)

Dataset: Multi-Session Chat (MSC)

- multi-session chat logs generated by human labelers, each playing a consistent persona
- MSC has five total sessions, and each session consists of a roughly a dozen messages

Persona A:
I'm married to my wife and I have 2 kids.
I haven't been to Disneyland.
I love going to concerts. Rolling Stones are one of my favourite bands.
I have an economics degree.
I recently started a new job that values my degree. I like my new job.

Persona B:
I'm close with my family.
I have been to Disneyland. I love Mickey Mouse.
I work as a baby delivery nurse.
I'd like to learn to coupon to save money. I applied for scholarships.

A: Hello, how are you doing?
B: I love spending time with my family.
A: That is great, me too! I'm married and my husband and I have 2 children.
B: So then have you ever been to Disneyland?
A: No, we recently purchased a new house, so we cannot afford it. Have you?
B: Yes, I love Mickey Mouse — such a cute little rat!
A: I enjoy going to concerts; I see The Rolling Stones every year.
....

---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for conversational agents (recall memory and working memory)

Dataset: Multi-Session Chat (MSC)

- multi-session chat logs generated by human labelers, each playing a consistent persona
- MSC has five total sessions, and each session consists of a roughly a dozen messages
- For consistency experiments, they created a new session (session 6) that contains a single question - answer response pair between the same two personas

B: Hey, remember that time we talked about music? What was the artist you mentioned you could get into?
A: Taylor Swift!

B: Hey, remember that time we talked about our jobs and expenses? What was that one thing you said you did to save money?
A: I eat a fresh and raw diet to save on groceries.

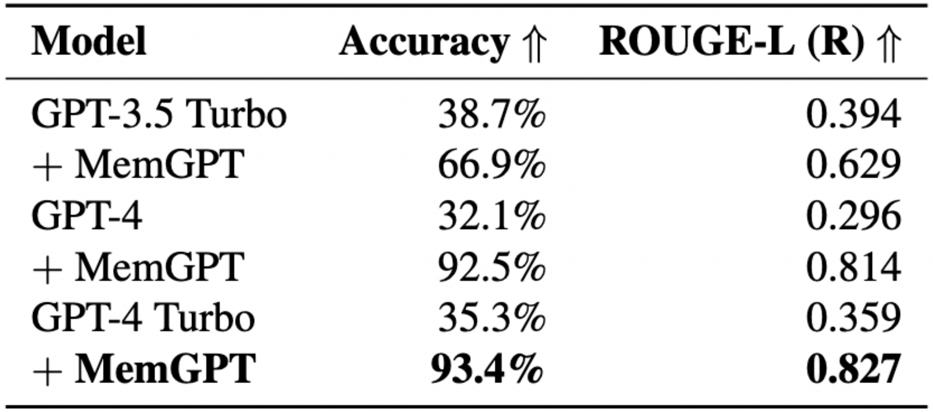
---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for conversational agents (recall memory and working memory)

Consistency: agent is asked a question by the user that explicitly refers to a prior conversation and has a very narrow expected answer range



Model	**Accuracy ⇑**	**ROUGE-L (R) ⇑**
GPT-3.5 Turbo	38.7%	0.394
+ MemGPT	66.9%	0.629
GPT-4	32.1%	0.296
+ MemGPT	92.5%	0.814
GPT-4 Turbo	35.3%	0.359
+ MemGPT	**93.4%**	**0.827**

Consistency

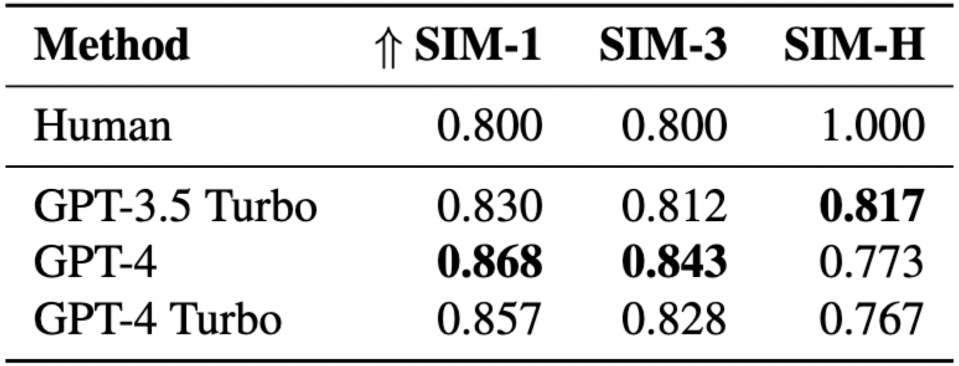
---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for conversational agents (recall memory and working memory)

Engagement: agent’s ability to craft engaging messages to the user that draw from knowledge accumulated in prior conversations.



Method	⇑ SIM-1	SIM-3	SIM-H
Human	0.800	0.800	1.000
GPT-3.5 Turbo	0.830	0.812	**0.817**
GPT-4	**0.868**	**0.843**	0.773
GPT-4 Turbo	0.857	0.828	0.767

Engagement

---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

1. MemGPT for conversational agents (recall memory and working memory)
   1. Consistency - agent should maintain conversational coherence
   2. Engagement - agent should draw on long-term knowledge about the user to personalize

2. **MemGPT for document analysis (archival memory)**
   1. Ability to infer from multiple documents

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for document analysis (archival memory)

Dataset: NaturalQuestions-Open dataset

- a retriever (`text-embedding-ada-002`) selects relevant Wikipedia documents for the questions

Q: Who was the ruler of England in 1616?
A: James I

Q: What is the Hot Coffee mod in *San Andreas*?
A: A normally inaccessible mini-game

Q: What is the maximum data rate for the 802.11a standard (select one)?
A: 54 Mbit/s

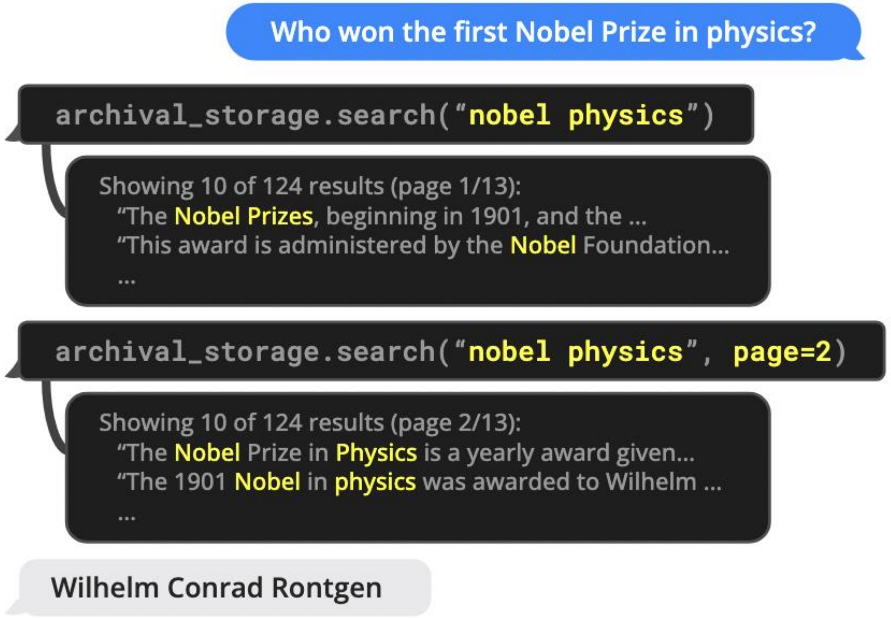
Q: Which state is located in the centre of India?
A: Chhattisgarh / Madhya Pradesh

---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for document analysis (archival memory)



Who won the first Nobel Prize in physics?

```
archival_storage.search("nobel physics")
```

Showing 10 of 124 results (page 1/13):
"The Nobel Prizes, beginning in 1901, and the ...
"This award is administered by the Nobel Foundation...
...

```
archival_storage.search("nobel physics", page=2)
```

Showing 10 of 124 results (page 2/13):
"The Nobel Prize in Physics is a yearly award given...
"The 1901 Nobel in physics was awarded to Wilhelm ...
...

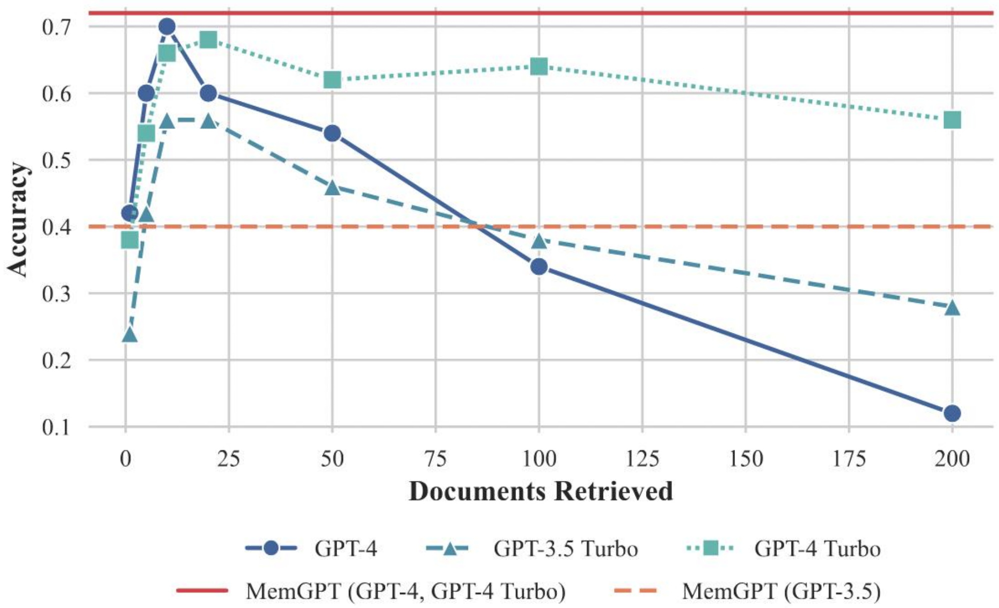
Wilhelm Conrad Rontgen

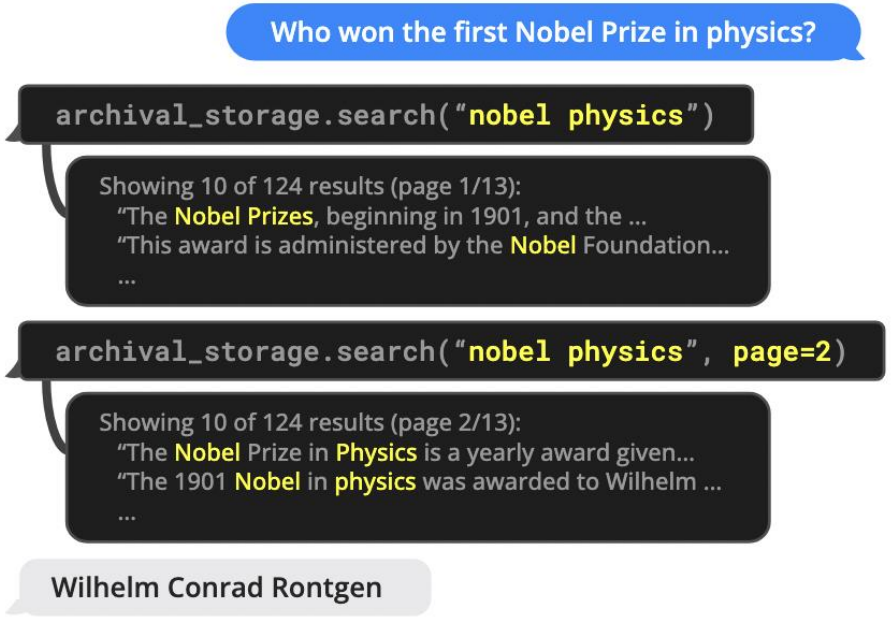
---

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

MemGPT: Towards LLMs as Operating Systems *

MemGPT for document analysis (archival memory)





Who won the first Nobel Prize in physics?

```
archival_storage.search("nobel physics")
```

Showing 10 of 124 results (page 1/13):
"The Nobel Prizes, beginning in 1901, and the ...
"This award is administered by the Nobel Foundation...
...

```
archival_storage.search("nobel physics", page=2)
```

Showing 10 of 124 results (page 2/13):
"The Nobel Prize in Physics is a yearly award given...
"The 1901 Nobel in physics was awarded to Wilhelm ...
...

Wilhelm Conrad Rontgen

* MemGPT: Towards LLMs as Operating Systems, Packer et al, Feb 2024

Outline

- Memory Management in AI Agents
- **How can a small-sized LLM approach larger proprietary model performance?**

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

- Introduces a new benchmark named `ColBench`
    - an LLM agent interacts with a human collaborator over multiple turns to solve realistic tasks in backend programming and frontend design

- proposes a novel RL algorithm, **`SWEET-RL`** (**`RL`** with **`S`**tep-**`W`**is**`E`** **`E`**valuation from **`T`**raining-time information)
    - trains a critic model with access to additional training-time information
    - critic provides step-level rewards for improving the policy mode

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

ColBench: *(1) Backend Programming Collaborations*

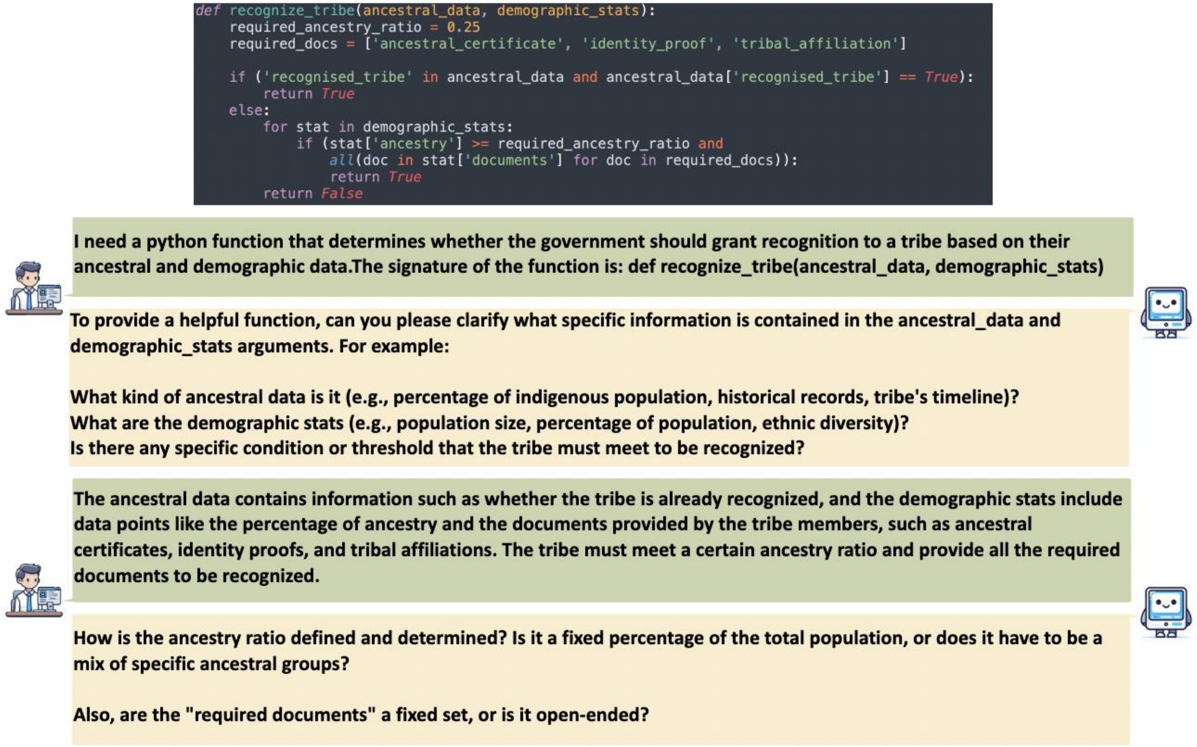
- Task:
  - Agent is required to collaborate with the human simulator to write a custom Python function (up to 50 lines)
  - In the first utterance, the agent is provided with a high-level description and the signature of the function
  - Agent has to reason and decide what clarifications are needed from the human simulator

- Dataset:
  - Python functions, high-level descriptions, and unit tests are generated by prompting Llama-3.1-70B-Instruct
  - Removed tasks where the generated python functions cannot pass their corresponding unit tests
  - Generated 10k such tasks in the train set and 1k tasks in the test set
  - 15k offline train trajectories are generated by zero-shot prompting
    - Llama-3.1-8B-Instruct as agent
    - Llama-3.1-70B-Instruct as human simulator

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

ColBench: *(1) Backend Programming Collaborations*



```python
def recognize_tribe(ancestral_data, demographic_stats):
    required_ancestry_ratio = 0.25
    required_docs = ['ancestral_certificate', 'identity_proof', 'tribal_affiliation']

    if ('recognised_tribe' in ancestral_data and ancestral_data['recognised_tribe'] == True):
        return True
    else:
        for stat in demographic_stats:
            if (stat['ancestry'] >= required_ancestry_ratio and
                all(doc in stat['documents'] for doc in required_docs)):
                return True
        return False
```

I need a python function that determines whether the government should grant recognition to a tribe based on their ancestral and demographic data.The signature of the function is: def recognize_tribe(ancestral_data, demographic_stats)

To provide a helpful function, can you please clarify what specific information is contained in the ancestral_data and demographic_stats arguments. For example:

What kind of ancestral data is it (e.g., percentage of indigenous population, historical records, tribe's timeline)?
What are the demographic stats (e.g., population size, percentage of population, ethnic diversity)?
Is there any specific condition or threshold that the tribe must meet to be recognized?

The ancestral data contains information such as whether the tribe is already recognized, and the demographic stats include data points like the percentage of ancestry and the documents provided by the tribe members, such as ancestral certificates, identity proofs, and tribal affiliations. The tribe must meet a certain ancestry ratio and provide all the required documents to be recognized.

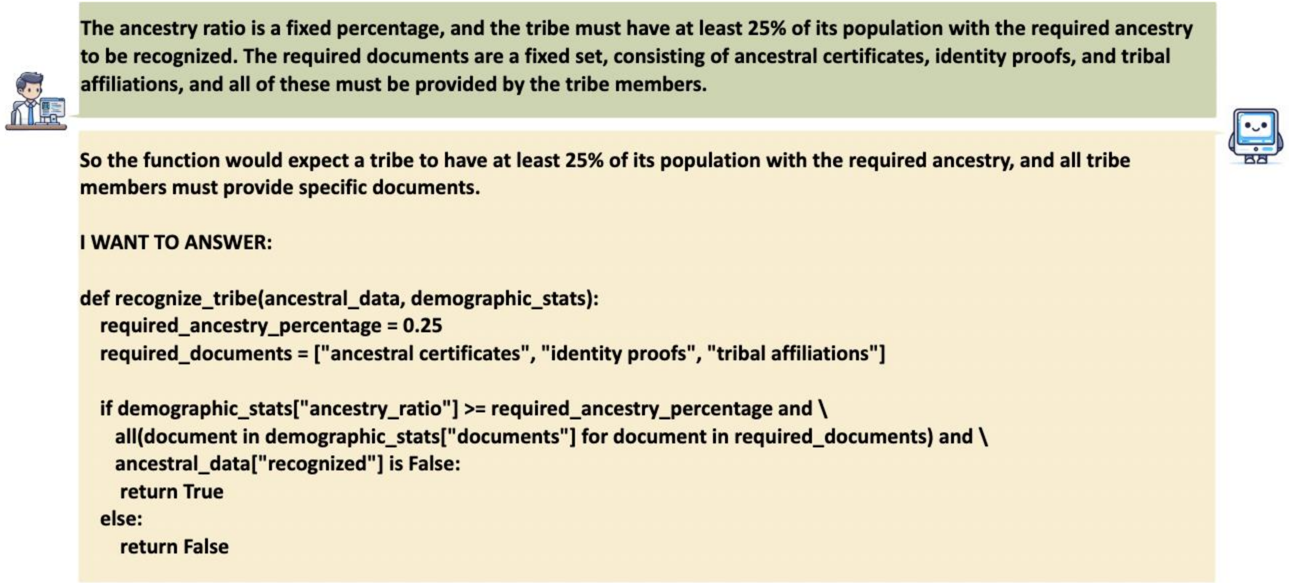
How is the ancestry ratio defined and determined? Is it a fixed percentage of the total population, or does it have to be a mix of specific ancestral groups?

Also, are the "required documents" a fixed set, or is it open-ended?

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

ColBench: *(1) Backend Programming Collaborations*



The ancestry ratio is a fixed percentage, and the tribe must have at least 25% of its population with the required ancestry to be recognized. The required documents are a fixed set, consisting of ancestral certificates, identity proofs, and tribal affiliations, and all of these must be provided by the tribe members.

So the function would expect a tribe to have at least 25% of its population with the required ancestry, and all tribe members must provide specific documents.

I WANT TO ANSWER:

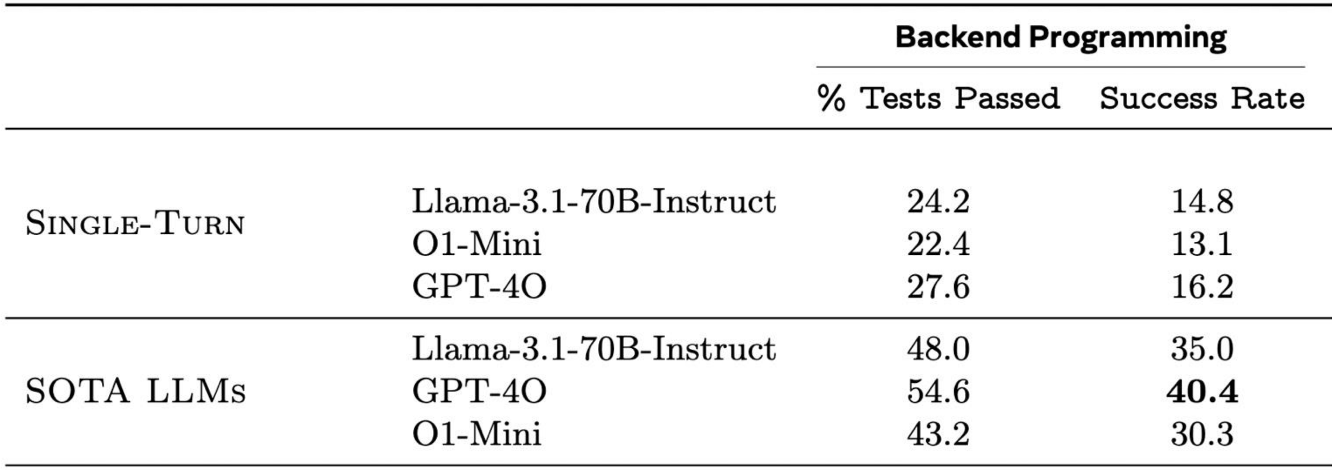
```
def recognize_tribe(ancestral_data, demographic_stats):
   required_ancestry_percentage = 0.25
   required_documents = ["ancestral certificates", "identity proofs", "tribal affiliations"]

   if demographic_stats["ancestry_ratio"] >= required_ancestry_percentage and \
     all(document in demographic_stats["documents"] for document in required_documents) and \
     ancestral_data["recognized"] is False:
      return True
   else:
      return False
```

Unit Tests

---

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*



		Backend Programming	
		% Tests Passed	Success Rate
SINGLE-TURN	Llama-3.1-70B-Instruct	24.2	14.8
	O1-Mini	22.4	13.1
	GPT-4O	27.6	16.2
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0
	GPT-4O	54.6	**40.4**
	O1-Mini	43.2	30.3

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

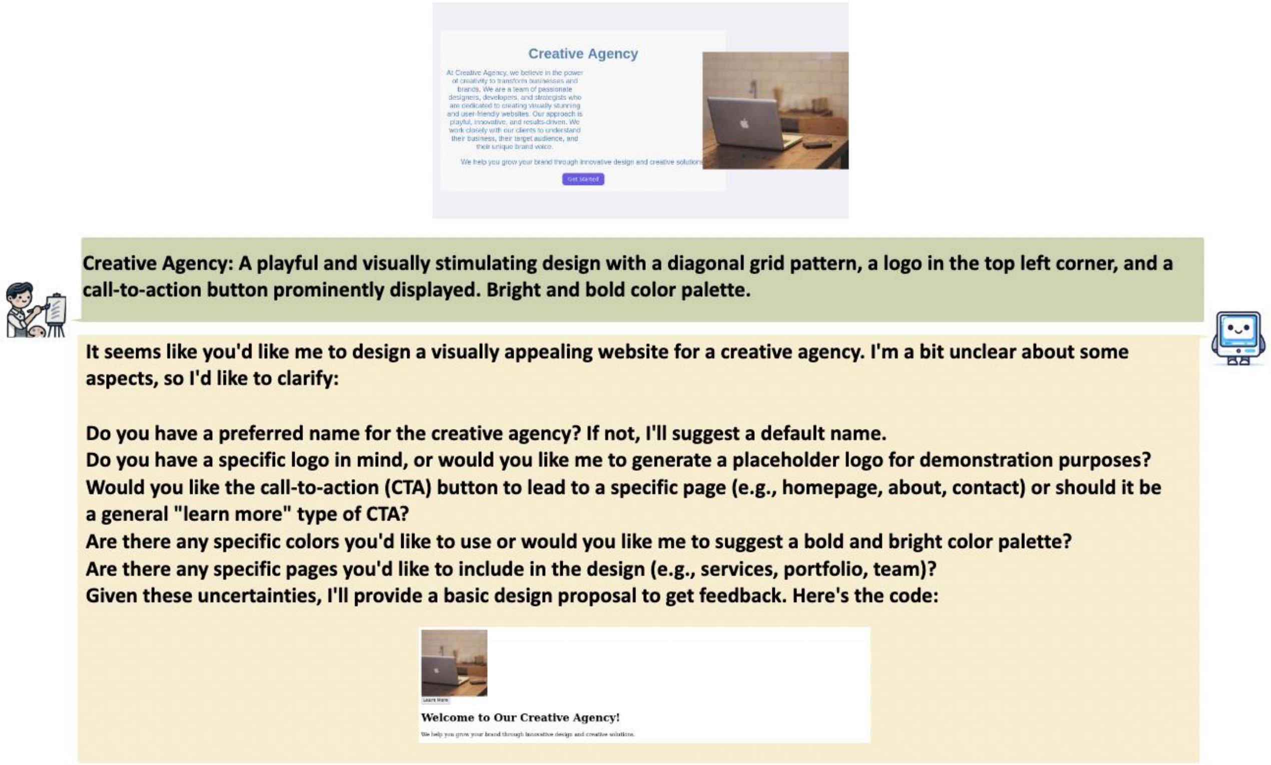
ColBench: *(2) Frontend Design Collaborations*

- Task:
  - agent is required to collaborate with the human simulator to design a web page by writing an HTML snippet (around 100 lines)
  - in the first utterance, the agent is provided with a high-level description of the web page
  - details such as the layout and color palette of the web page are missing and only visible to the human simulator
- Dataset:
  - reference web pages and high-level descriptions are from WebSight (Laurençon et al., 2024)
  - 10k such tasks for training and 500 for the test set
  - 6k offline train trajectories are generated by zero-shot prompting
    - Llama-3.1-8B-Instruct as agent
    - Qwen2-VL-72B as human simulator.

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

ColBench: *(2) Frontend Design Collaborations*



Creative Agency

At Creative Agency, we believe in the power of creativity to transform businesses and brands. We are a team of passionate designers, developers, and strategists who are dedicated to creating visually stunning and user-friendly websites. Our approach is playful, innovative, and results-driven. We work closely with our clients to understand their business, their target audience, and their unique brand voice.

We help you grow your brand through innovative design and creative solutions

Creative Agency: A playful and visually stimulating design with a diagonal grid pattern, a logo in the top left corner, and a call-to-action button prominently displayed. Bright and bold color palette.

It seems like you'd like me to design a visually appealing website for a creative agency. I'm a bit unclear about some aspects, so I'd like to clarify:

Do you have a preferred name for the creative agency? If not, I'll suggest a default name.
Do you have a specific logo in mind, or would you like me to generate a placeholder logo for demonstration purposes?
Would you like the call-to-action (CTA) button to lead to a specific page (e.g., homepage, about, contact) or should it be a general "learn more" type of CTA?
Are there any specific colors you'd like to use or would you like me to suggest a bold and bright color palette?
Are there any specific pages you'd like to include in the design (e.g., services, portfolio, team)?
Given these uncertainties, I'll provide a basic design proposal to get feedback. Here's the code:

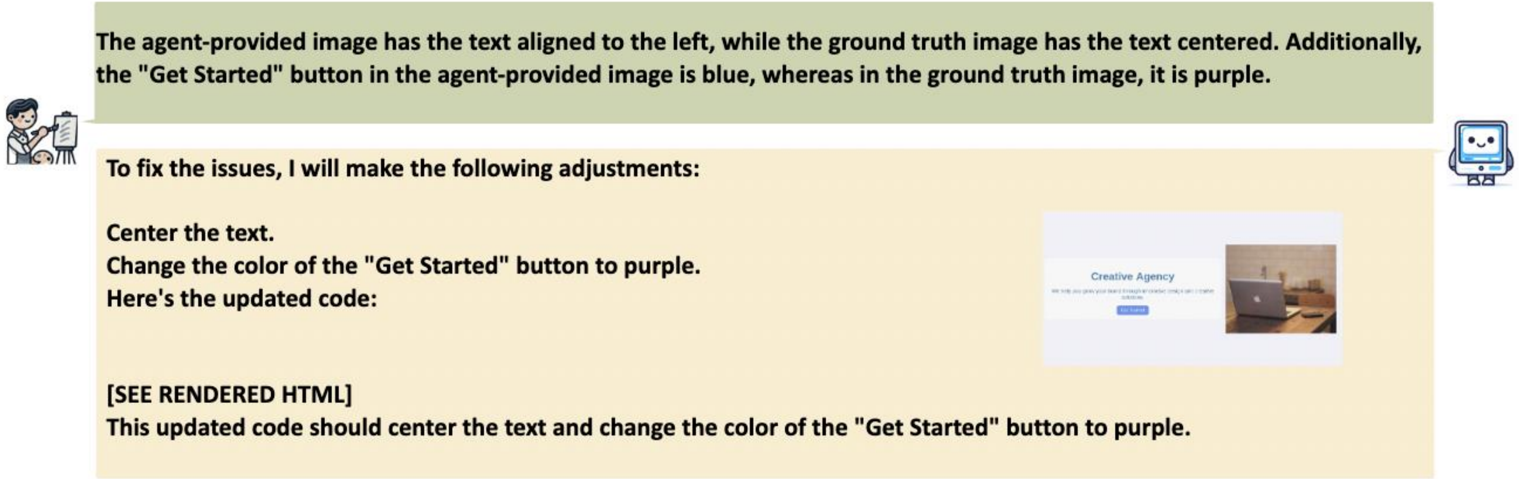
Welcome to Our Creative Agency!

We help you grow your brand through innovative design and creative solutions.

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

ColBench: *(2) Frontend Design Collaborations*



The agent-provided image has the text aligned to the left, while the ground truth image has the text centered. Additionally, the "Get Started" button in the agent-provided image is blue, whereas in the ground truth image, it is purple.

To fix the issues, I will make the following adjustments:

Center the text.
Change the color of the "Get Started" button to purple.
Here's the updated code:

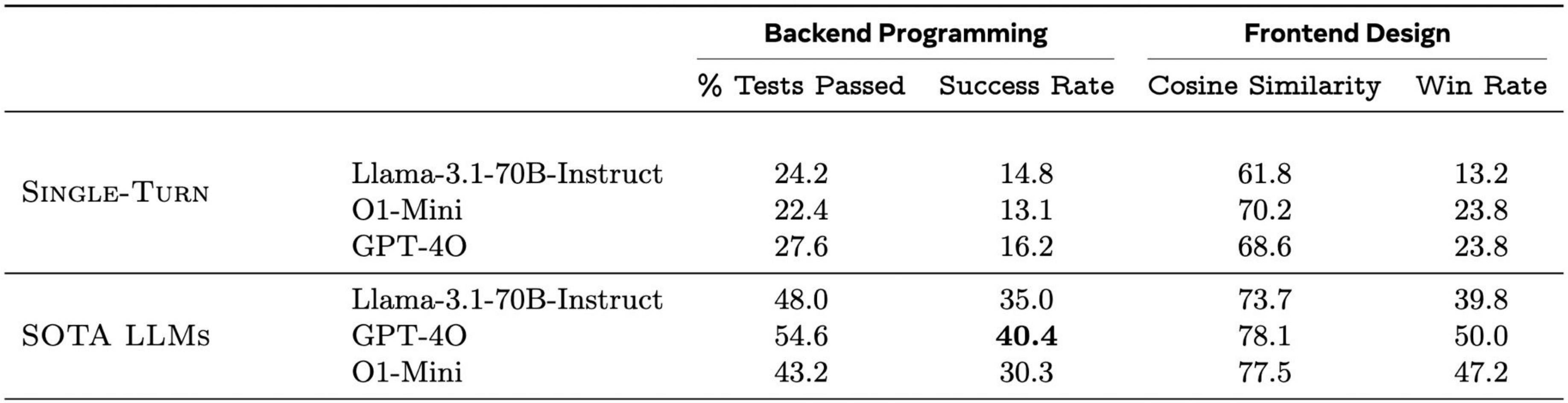
[SEE RENDERED HTML]
This updated code should center the text and change the color of the "Get Started" button to purple.

Image Similarity

0.883

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

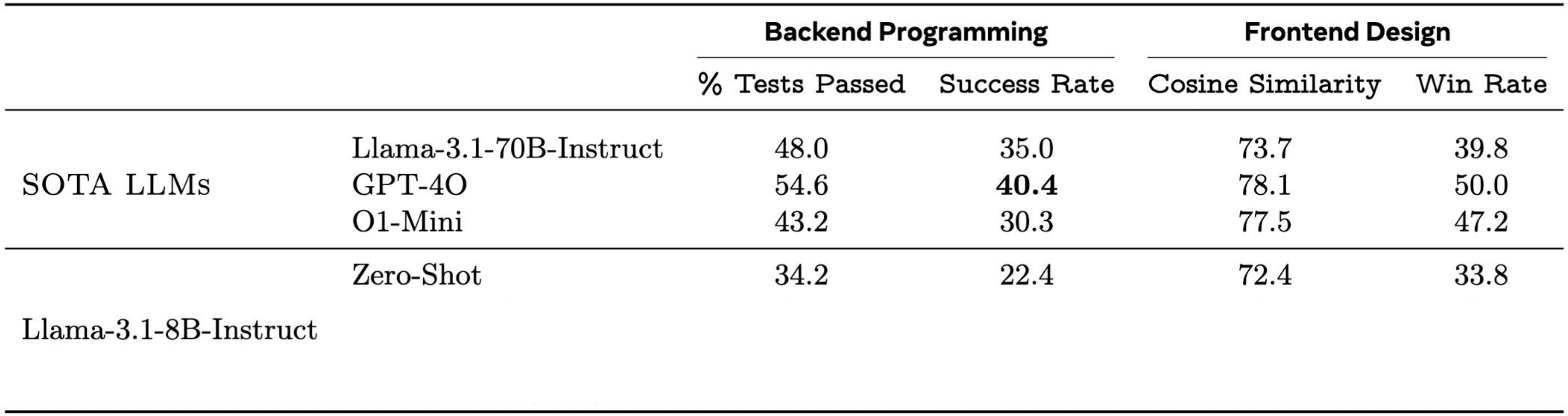


		Backend Programming		**Frontend Design**	
		% Tests Passed	Success Rate	Cosine Similarity	Win Rate
SINGLE-TURN	Llama-3.1-70B-Instruct	24.2	14.8	61.8	13.2
	O1-Mini	22.4	13.1	70.2	23.8
	GPT-4O	27.6	16.2	68.6	23.8
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0	73.7	39.8
	GPT-4O	54.6	**40.4**	78.1	50.0
	O1-Mini	43.2	30.3	77.5	47.2

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

How can a small-sized LLM approach larger proprietary model performance?

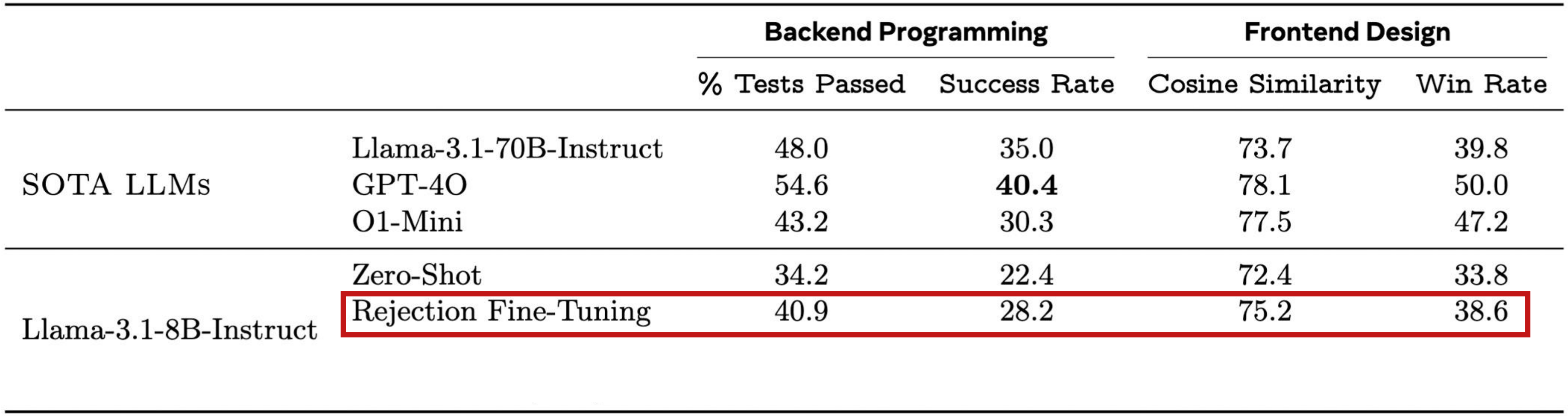


		Backend Programming		Frontend Design	
		% Tests Passed	Success Rate	Cosine Similarity	Win Rate
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0	73.7	39.8
	GPT-4O	54.6	**40.4**	78.1	50.0
	O1-Mini	43.2	30.3	77.5	47.2
Llama-3.1-8B-Instruct	Zero-Shot	34.2	22.4	72.4	33.8

---

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*



		Backend Programming		**Frontend Design**	
		% Tests Passed	Success Rate	Cosine Similarity	Win Rate
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0	73.7	39.8
	GPT-4O	54.6	**40.4**	78.1	50.0
	O1-Mini	43.2	30.3	77.5	47.2
Llama-3.1-8B-Instruct	Zero-Shot	34.2	22.4	72.4	33.8
	Rejection Fine-Tuning	40.9	28.2	75.2	38.6

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

RECAP

Positive Trajectory

$$\tau_{\mathrm{w}} = (\mathrm{u}_1^{\mathrm{w}}, \mathrm{a}_1^{\mathrm{w}}, \mathrm{t}_1^{\mathrm{w}}, \cdots, \mathrm{u}_{\mathrm{H_w}}^{\mathrm{w}}, \mathrm{a}_{\mathrm{H_w}}^{\mathrm{w}}, \mathrm{t}_{\mathrm{H_w}}^{\mathrm{w}})$$

Negative Trajectory

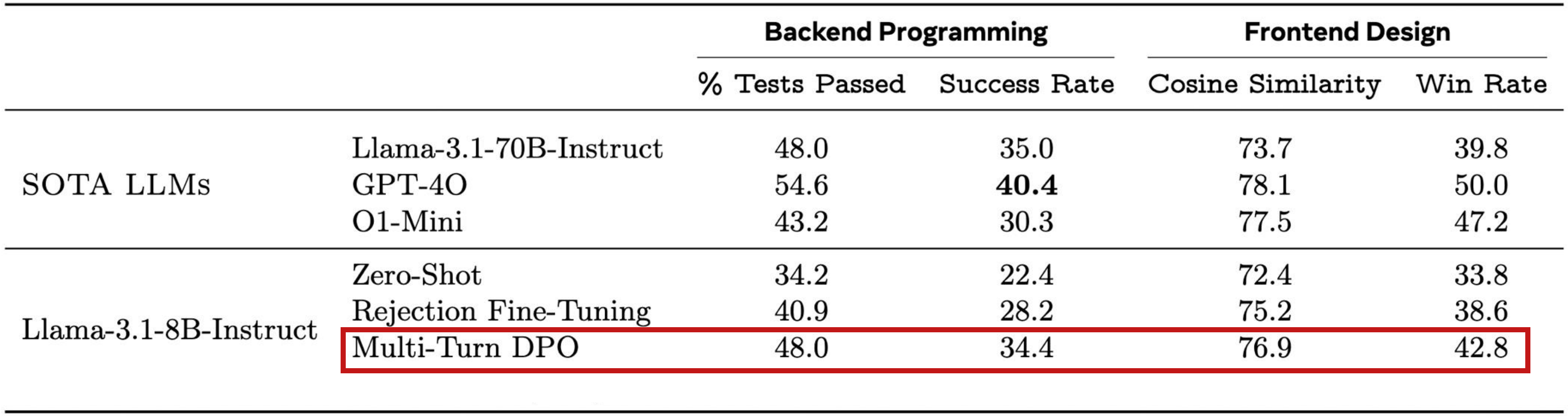
$$\tau_{\mathrm{l}} = (\mathrm{u}_1^{\mathrm{l}}, \mathrm{a}_1^{\mathrm{l}}, \mathrm{t}_1^{\mathrm{l}}, \cdots, \mathrm{u}_{\mathrm{H_l}}^{\mathrm{l}}, \mathrm{a}_{\mathrm{H_l}}^{\mathrm{l}}, \mathrm{t}_{\mathrm{H_l}}^{\mathrm{l}})$$

$$\mathcal{L}(x; \tau_w, \tau_l) = \mathcal{L}_{\mathrm{SFT}}(x; \tau_w) + \lambda \mathcal{L}_{\mathrm{mDPO}}(x; \tau_w, \tau_l),$$

$$\mathcal{L}_{\mathrm{mDPO}}(x; \tau_w, \tau_l) = -\log\sigma\left(\eta\left(\sum_{\tau_l} \frac{\pi_\theta(a^l|s^l)}{\pi_{ref}(a^l|s^l)} - \sum_{\tau_w} \frac{\pi_\theta(a^w|s^w)}{\pi_{ref}(a^w|s^w)}\right)\right)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*



		Backend Programming		Frontend Design	
		% Tests Passed	Success Rate	Cosine Similarity	Win Rate
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0	73.7	39.8
	GPT-4O	54.6	**40.4**	78.1	50.0
	O1-Mini	43.2	30.3	77.5	47.2
Llama-3.1-8B-Instruct	Zero-Shot	34.2	22.4	72.4	33.8
	Rejection Fine-Tuning	40.9	28.2	75.2	38.6
	Multi-Turn DPO	48.0	34.4	76.9	42.8

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

RECAP

Positive Trajectory

$$\tau_{\mathrm{w}} = (\mathrm{u}_1^{\mathrm{w}}, \mathrm{a}_1^{\mathrm{w}}, \mathrm{t}_1^{\mathrm{w}}, \cdots, \mathrm{u}_{\mathrm{H}_{\mathrm{w}}}^{\mathrm{w}}, \mathrm{a}_{\mathrm{H}_{\mathrm{w}}}^{\mathrm{w}}, \mathrm{t}_{\mathrm{H}_{\mathrm{w}}}^{\mathrm{w}})$$

Negative Trajectory

$$\tau_{\mathrm{l}} = (\mathrm{u}_1^{\mathrm{l}}, \mathrm{a}_1^{\mathrm{l}}, \mathrm{t}_1^{\mathrm{l}}, \cdots, \mathrm{u}_{\mathrm{H}_{\mathrm{l}}}^{\mathrm{l}}, \mathrm{a}_{\mathrm{H}_{\mathrm{l}}}^{\mathrm{l}}, \mathrm{t}_{\mathrm{H}_{\mathrm{l}}}^{\mathrm{l}})$$

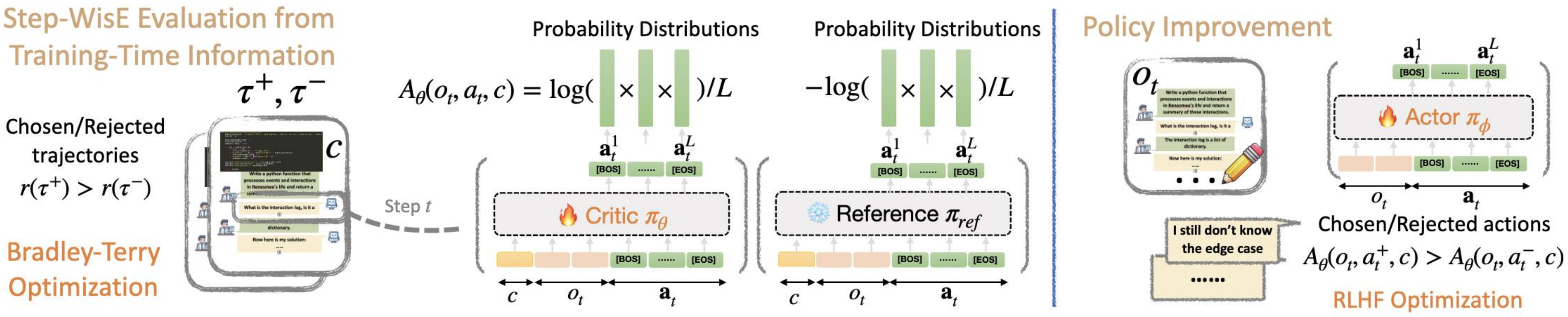
Do not perform explicit credit assignment across turns ☹

$$\mathcal{L}(x; \tau_w, \tau_l) = \mathcal{L}_{\mathrm{SFT}}(x; \tau_w) + \lambda \mathcal{L}_{\mathrm{mDPO}}(x; \tau_w, \tau_l),$$

$$\mathcal{L}_{\mathrm{mDPO}}(x; \tau_w, \tau_l) = -\log\sigma\left(\eta\left(\sum_{\tau_l} \frac{\pi_\theta(a^l|s^l)}{\pi_{ref}(a^l|s^l)} - \sum_{\tau_w} \frac{\pi_\theta(a^w|s^w)}{\pi_{ref}(a^w|s^w)}\right)\right)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*


Step-WisE Evaluation from Training-Time Information
τ^+, τ^-
Chosen/Rejected trajectories
$r(\tau^+) > r(\tau^-)$
c
Bradley-Terry Optimization
Step t
Probability Distributions
$A_\theta(o_t, a_t, c) = \log(\times \times)/L$
$\mathbf{a}_t^1$
$\mathbf{a}_t^L$
[BOS]
......
[EOS]
Critic π_θ
c
o_t
$\mathbf{a}_t$
Probability Distributions
$-\log(\times \times)/L$
Reference π_{ref}
Policy Improvement
O_t
Actor π_ϕ
I still don't know the edge case
......
Chosen/Rejected actions
$A_\theta(o_t, a_t^+, c) > A_\theta(o_t, a_t^-, c)$
RLHF Optimization

SWEET-RL:

1. Critic π_θ: Train a step-wise advantage function with **access to additional training-time information**
2. Actor π_ϕ : Improve policy by using the advantage function as a reward model

---

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

The problem of collaboration between humans and agents is framed as an POMDP $\mathcal{M} = \{O, C, A, T, \mu_1, R, N\}$

- O and C are the observable and hidden parts of the state space

- $o_1 \in O$ is the initial user utterance that is observable
- o_t is the entire interaction history at time t

- $\mathrm{c} \in C$ is the hidden part of the state, only available during train
    - test cases (backend programming)
    - reference webpage of the specification (frontend design)

- $a_\mathrm{t} \in A$ is the agent response with L tokens $a_\mathrm{t}^{1:\mathrm{L}}$

- At each step, the agent receives a scalar reward $r(o_t, a_t, c) \in R$

- objective is to maximize the cumulative reward $\sum_{t=1}^{N} r(o_t, a_t, c)$, where N is the max allowed turns

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

Q-function for a policy π represents the expected cumulative reward starting from turn t :

$$\mathrm{Q}^{\pi}(\mathrm{o}_{\mathrm{t}}, a_{\mathrm{t}}, \mathrm{c}) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{N} r(o_{t'}, a_{t'}, c)\right]$$

The value function is the expected Q-value under the policy's action distribution:

$$\mathrm{V}^{\pi}(\mathrm{o}_{\mathrm{t}}, \mathrm{c}) = \mathbb{E}_{a_{\mathrm{t}}\sim\pi}[\mathrm{Q}^{\pi}(\mathrm{o}_{\mathrm{t}}, a_{\mathrm{t}}, \mathrm{c})]$$

The advantage function measures how much better an action a_{t} is compared to the average action under policy π:

$$\mathrm{A}^{\pi}(\mathrm{o}_{\mathrm{t}}, a_{\mathrm{t}}, \mathrm{c}) = \mathrm{Q}^{\pi}(\mathrm{o}_{\mathrm{t}}, a_{\mathrm{t}}, \mathrm{c}) - \mathrm{V}^{\pi}(\mathrm{o}_{\mathrm{t}}, \mathrm{c})$$

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

Let τ^+ and τ^- be the chosen and rejected trajectories under the same task and same additional training information

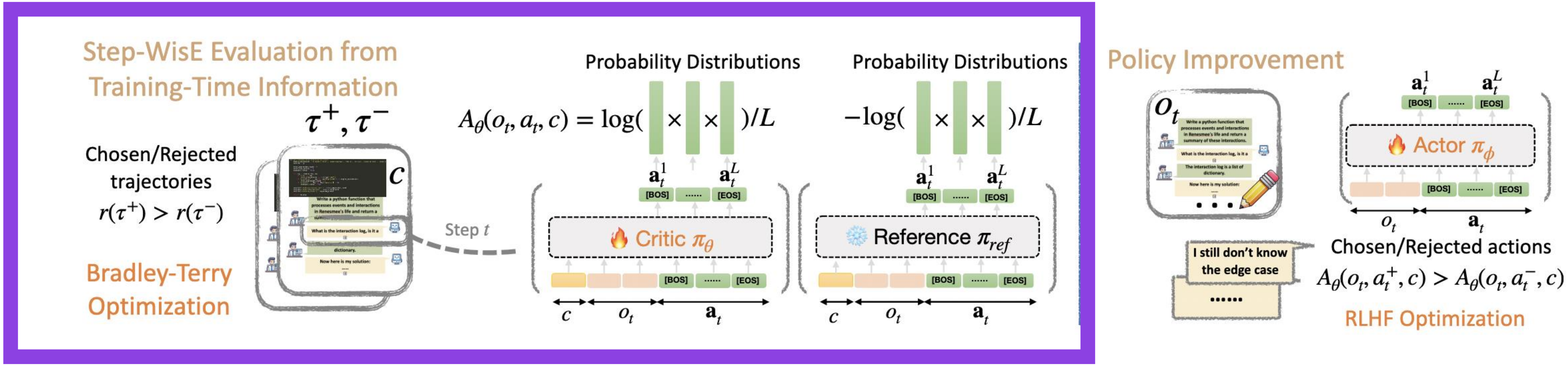
$$\mathcal{J}_{\mathrm{BT}} = -\log\left[\sigma\left(\sum_t \beta r(o_t^+, a_t^+, c) - \sum_t \beta r(o_t^-, a_t^-, c)\right)\right]$$

$$\mathcal{J}_A(\theta) = -\log\left[\sigma\left(\sum_t \beta A_\theta(o_t^+, a_t^+, c) - \sum_t \beta A_\theta(o_t^-, a_t^-, c)\right)\right]$$

$$A_\theta(o_t, a_t, h) = \frac{1}{L}\sum_{l=1}^{L}\left[\log \frac{\pi_\theta(a_t^l | o_t, a_t^{1:l-1}, c)}{\pi_{\mathrm{ref}}(a_t^l | o_t, a_t^{1:l-1}, c)}\right]$$

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*


Step-WisE Evaluation from Training-Time Information
τ^+, τ^-
Chosen/Rejected trajectories
$r(\tau^+) > r(\tau^-)$
Bradley-Terry Optimization
c
Step t
Probability Distributions
$A_\theta(o_t, a_t, c) = \log(\times \times)/L$
Probability Distributions
$-\log(\times \times)/L$
$\mathbf{a}_t^1$
$\mathbf{a}_t^L$
[BOS]
......
[EOS]
Critic π_θ
Reference π_{ref}
c
o_t
$\mathbf{a}_t$
Policy Improvement
o_t
Actor π_ϕ
I still don't know the edge case
......
Chosen/Rejected actions
$A_\theta(o_t, a_t^+, c) > A_\theta(o_t, a_t^-, c)$
RLHF Optimization

SWEET-RL:

1. Critic π_θ: Train a step-wise advantage function with access to additional training-time information
2. Actor π_ϕ : Improve policy by using the advantage function as a reward model

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*

For each turn, sample 16 candidate actions, compute their turn wise advantage A_θ

- take random action from top-50% quantile as chosen
- take random action from bottom-50% quantile as rejected

$$\mathcal{J}_\pi(\phi) = -\log\sigma\left(\beta'\frac{\log\pi_\phi(a^+|o_t)}{\log\pi_{\mathrm{ref}}(a^+|o_t)} - \beta'\frac{\log\pi_\phi(a^-|o_t)}{\log\pi_{\mathrm{ref}}(a^-|o_t)}\right)$$

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

mDPO vs SWEET-RL

Positive Trajectory

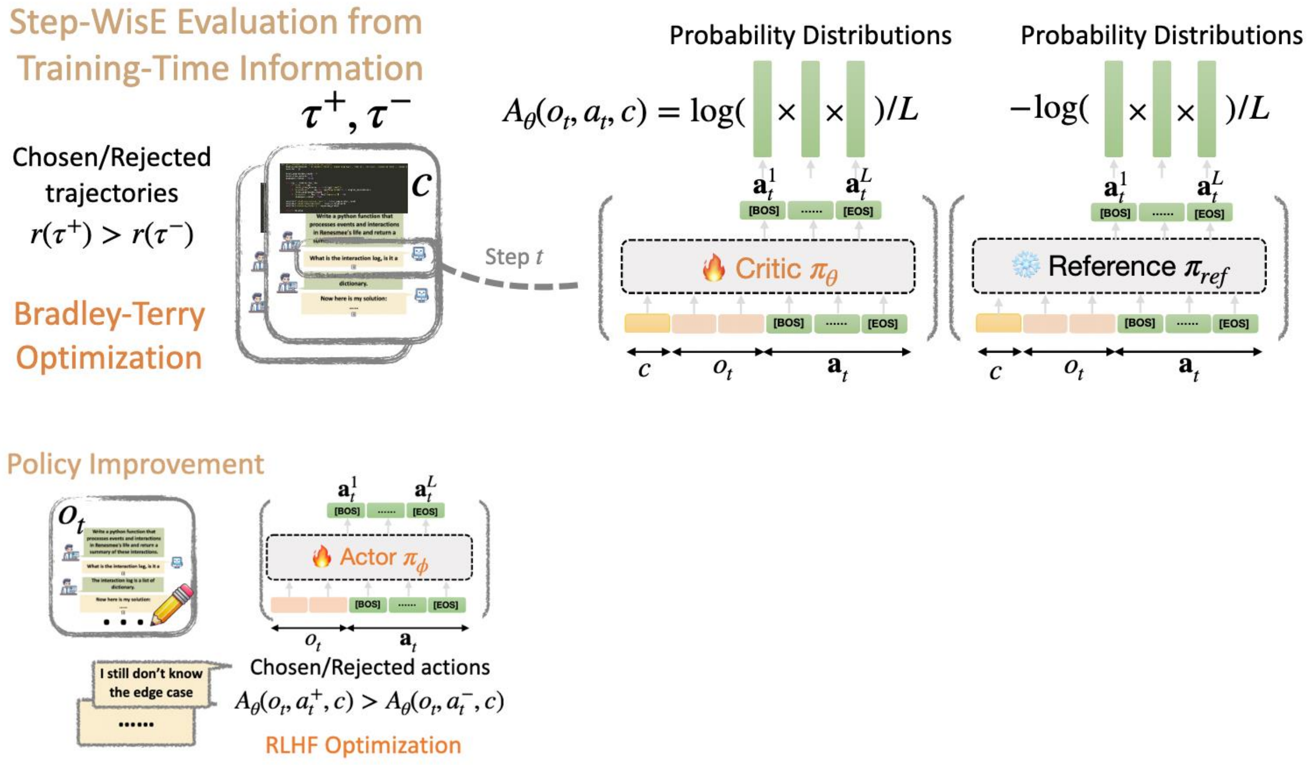
$$\tau_{\mathrm{w}} = (\mathrm{u}_1^{\mathrm{w}}, \mathrm{a}_1^{\mathrm{w}}, \mathrm{t}_1^{\mathrm{w}}, \cdots, \mathrm{u}_{\mathrm{H_w}}^{\mathrm{w}}, \mathrm{a}_{\mathrm{H_w}}^{\mathrm{w}}, \mathrm{t}_{\mathrm{H_w}}^{\mathrm{w}})$$

Negative Trajectory

$$\tau_{\mathrm{l}} = (\mathrm{u}_1^{\mathrm{l}}, \mathrm{a}_1^{\mathrm{l}}, \mathrm{t}_1^{\mathrm{l}}, \cdots, \mathrm{u}_{\mathrm{H_l}}^{\mathrm{l}}, \mathrm{a}_{\mathrm{H_l}}^{\mathrm{l}}, \mathrm{t}_{\mathrm{H_l}}^{\mathrm{l}})$$

$$\mathcal{L}(x;\tau_w,\tau_l) = \mathcal{L}_{\mathrm{SFT}}(x;\tau_w) + \lambda \mathcal{L}_{\mathrm{mDPO}}(x;\tau_w,\tau_l),$$

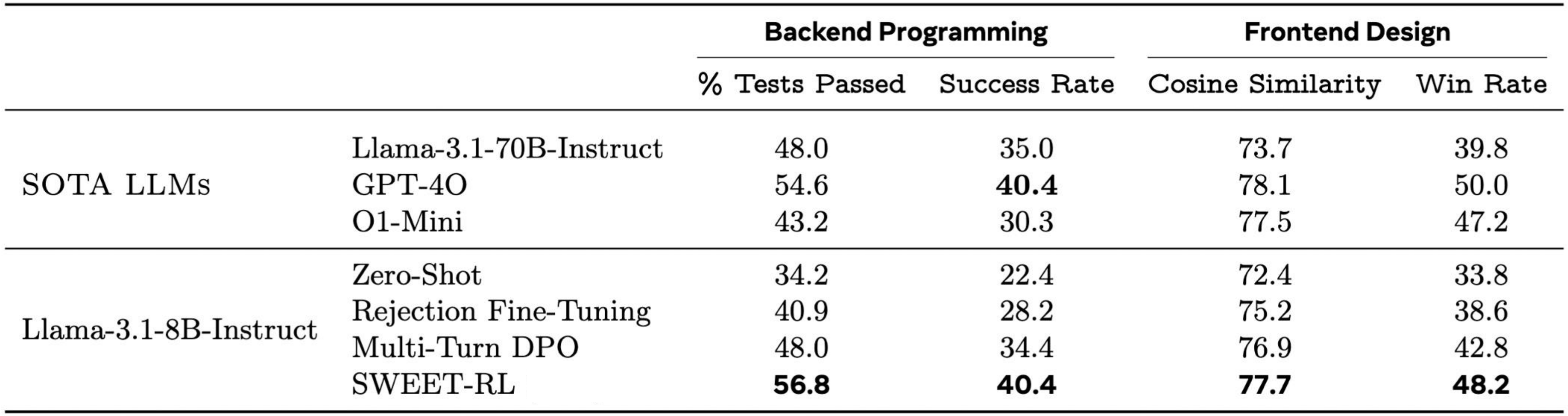
$$\mathcal{L}_{\mathrm{mDPO}}(x;\tau_w,\tau_l) = \text{-}\log\sigma\left(\eta\left(\sum_{\tau_l}\frac{\pi_\theta(a^l|s^l)}{\pi_{ref}(a^l|s^l)} - \sum_{\tau_w}\frac{\pi_\theta(a^w|s^w)}{\pi_{ref}(a^w|s^w)}\right)\right)$$


Step-WisE Evaluation from Training-Time Information
τ^+, τ^-
Chosen/Rejected trajectories
$r(\tau^+) > r(\tau^-)$
Bradley-Terry Optimization
c
Step t
Probability Distributions
Probability Distributions
$A_\theta(o_t, a_t, c) = \log(\ \times\ \times\ )/L$
$-\log(\ \times\ \times\ )/L$
$\mathbf{a}_t^1$
$\mathbf{a}_t^L$
[BOS] ...... [EOS]
Critic π_θ
Reference π_{ref}
[BOS] ...... [EOS]
c
o_t
$\mathbf{a}_t$
Policy Improvement
o_t
Actor π_ϕ
o_t
$\mathbf{a}_t$
Chosen/Rejected actions
$A_\theta(o_t, a_t^+, c) > A_\theta(o_t, a_t^-, c)$
RLHF Optimization
I still don't know the edge case
......

mDPO vs SWEET-RL

mDPO	SWEET-RL
No access to additional training time information	Has access to additional training time information
do not perform explicit credit assignment across turns	Uses a critic to improve turn-level credit assignment

SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks*



		Backend Programming		Frontend Design	
		% Tests Passed	Success Rate	Cosine Similarity	Win Rate
SOTA LLMs	Llama-3.1-70B-Instruct	48.0	35.0	73.7	39.8
	GPT-4O	54.6	**40.4**	78.1	50.0
	O1-Mini	43.2	30.3	77.5	47.2
Llama-3.1-8B-Instruct	Zero-Shot	34.2	22.4	72.4	33.8
	Rejection Fine-Tuning	40.9	28.2	75.2	38.6
	Multi-Turn DPO	48.0	34.4	76.9	42.8
	SWEET-RL	**56.8**	**40.4**	**77.7**	**48.2**

* SWEET-RL: Training Multi-Turn LLM Agents on Collaborative Reasoning Tasks, Zhou et al, Mar 2025

Summary

- Memory Management in AI Agents
  - MemGPT - an OS-inspired LLM system for virtual context management

- Small-sized LLM can approach larger proprietary model performance
  - Llama-3.1-8b trained with SWEET-RL approaches GPT4-O on ColBench