More Efficient NLP and Agents

吴家隆(Jialong Wu)
wujialongml@gmail.com

Overall

- Training Stage
  - **Less Parameter:** LoRA
  - **Less Data:** Active Learning (SATR)
    User Profiling (**PROPER**)
- Inference Stage
  - **Less Time:** Speculative Decoding (SEED)
  - **Less Memory:** KV-Cache Compression (**SCOPE**)
- Agent-Driven Autonomous Task Execution
  - **Less Human Intervention:**
    - **WebWalker**
    - **WebDancer**

Training Stage

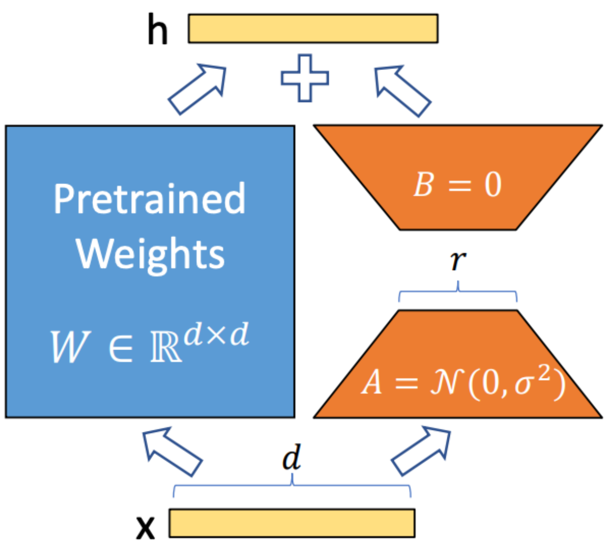


Figure 1: Low-rank adaptation.

- Finetuning Large Language Models (LLM) from scratch is quite resource-intensive, given the large number of parameters these models contain.
- Parameter efficient fine tuning is introduced.
- However, task-specific/user-specific fine-tuning also demands **high-quality data**.
- We propose **parameter and data efficient fine tuning**:
    - Active learning for task annotation
    - Progressive learning for personalized LLM

Training Stage

Active learning for task annotation

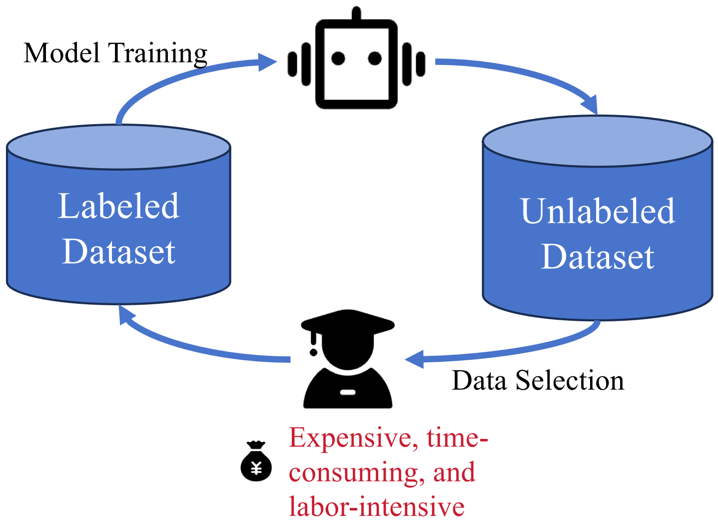


Figure 2: **Active Learning**.

Progressive learning for personalized LLM

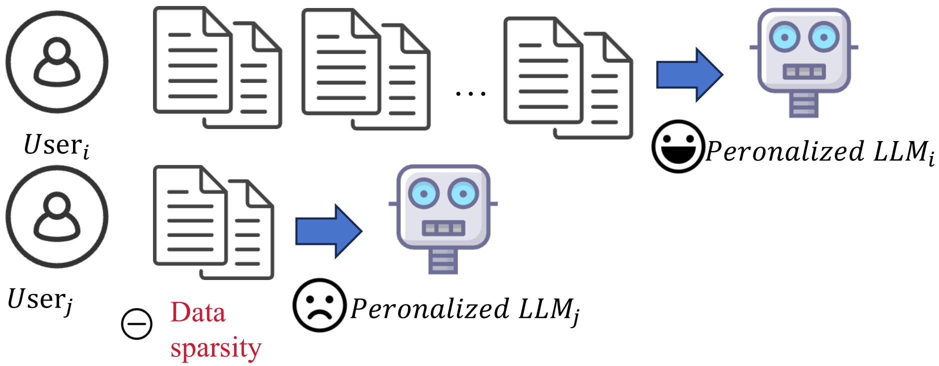


Figure 3: **Personalized LLM** Training.

Training Stage

STAR: Constraint LoRA with Dynamic Active Learning for Data-Efficient Fine-Tuning of Large Language Models

Probe experiments reveals:

- A clear **gap** between the base model and LoRA model
- Model **calibration** issue

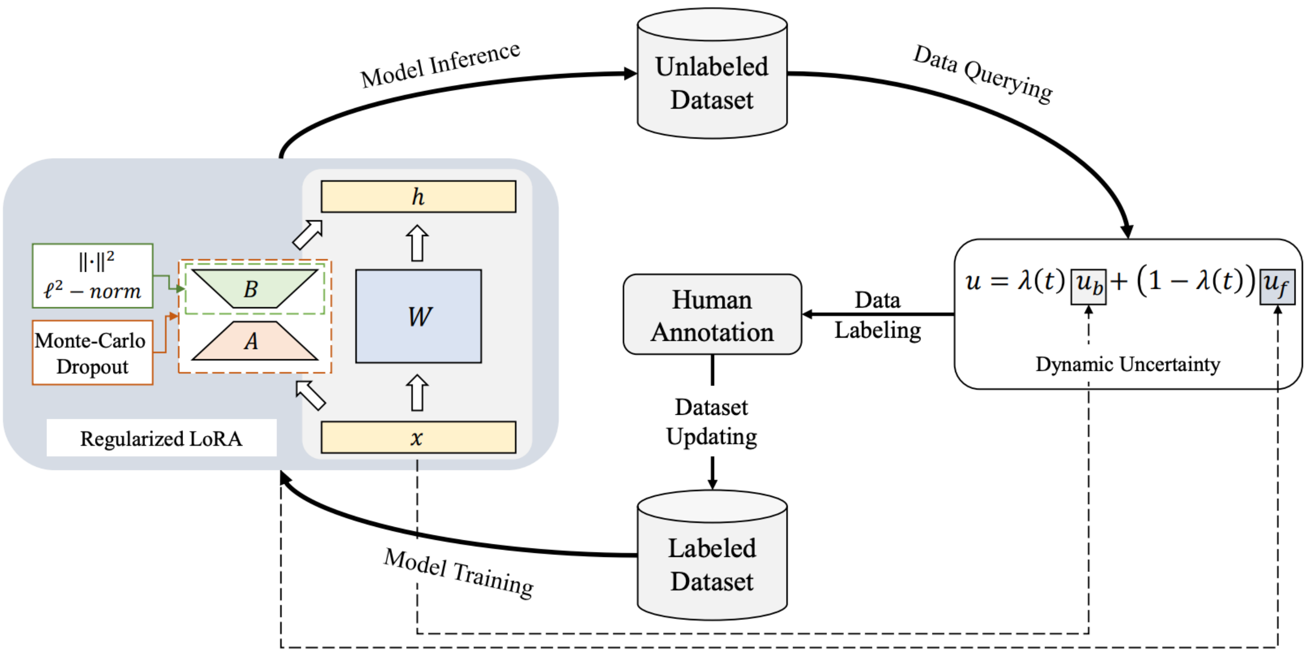


Figure 4: Overall of SATR.

Motivation: PROPER

- Current Large Language Models primarily offer a ***one-size-fits-all*** service.
- Personalized LLMs, tailoring the outputs to ***user-specific preferences***, became a hot research topic.
- Two types of LLM personalization methods: (1) **prompt-based**, (2) **fine-tuning-based**.

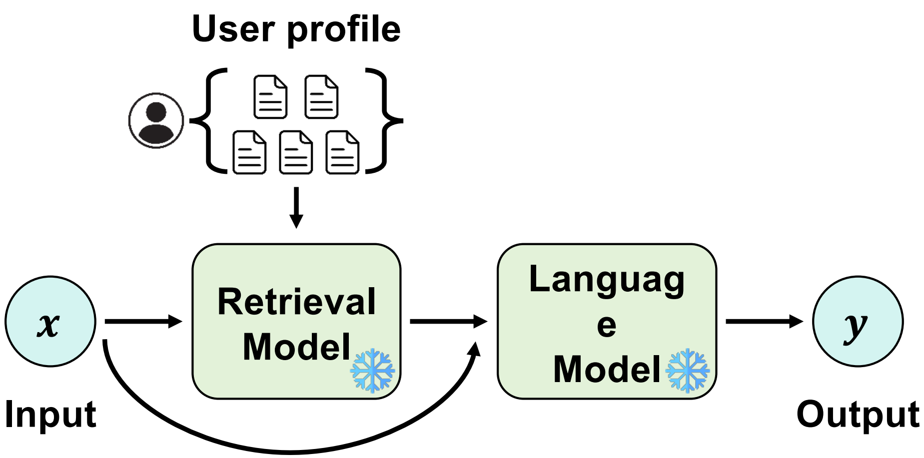


Fig. 1 (a) prompt-based personalized LLM,

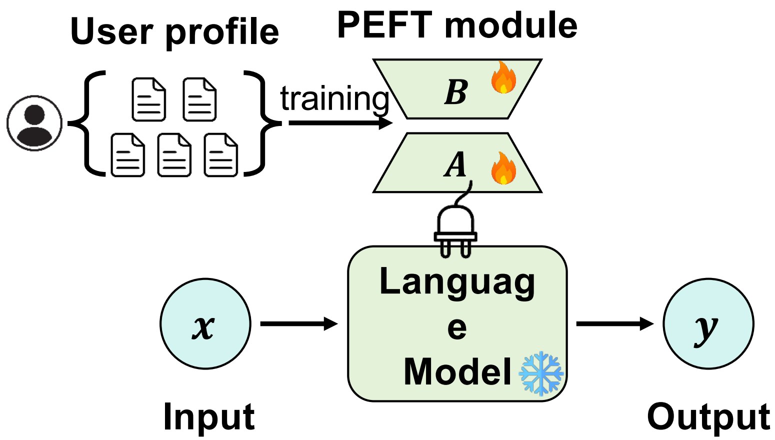


Fig. 1 (b) fine-tuning-based personalized LLM.

Fine-Tuning-based Personalized LLMs

- **Pros**: (1) data privacy, (2) better user behavior pattern generalization.
- **Cons**: (1) data scarcity for most of users (# avg tokens for task training is 20x of # avg tokens for individual users).
- **Solutions**: introduce meso-level LLMs (group-level LLMs) to bridge the macro-level LLMs (general LLMs) and micro-level LLMs (personalized LLMs).

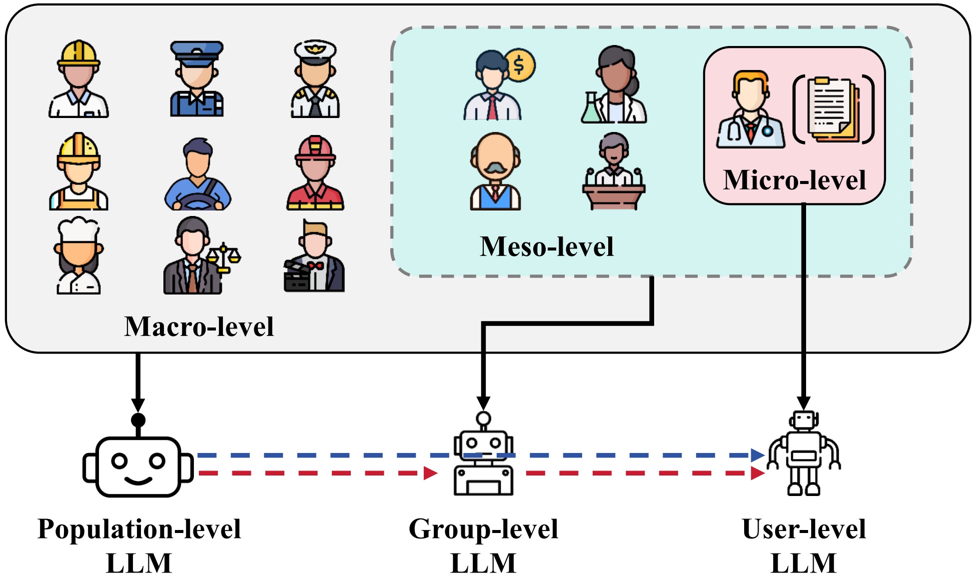


Fig 2. The comparison between different paradigms of LLM personalization.

Method

- A progressive learning framework, PROPER, consists of three stages: (1) population-level adaptation, (2) group-level adaptation, (3) user-level adaptation.
- Enable automatic user grouping via LoRAMoE and user-aware routers, while effectively integrating user and group-level knowledge through a LoRA-aware router.

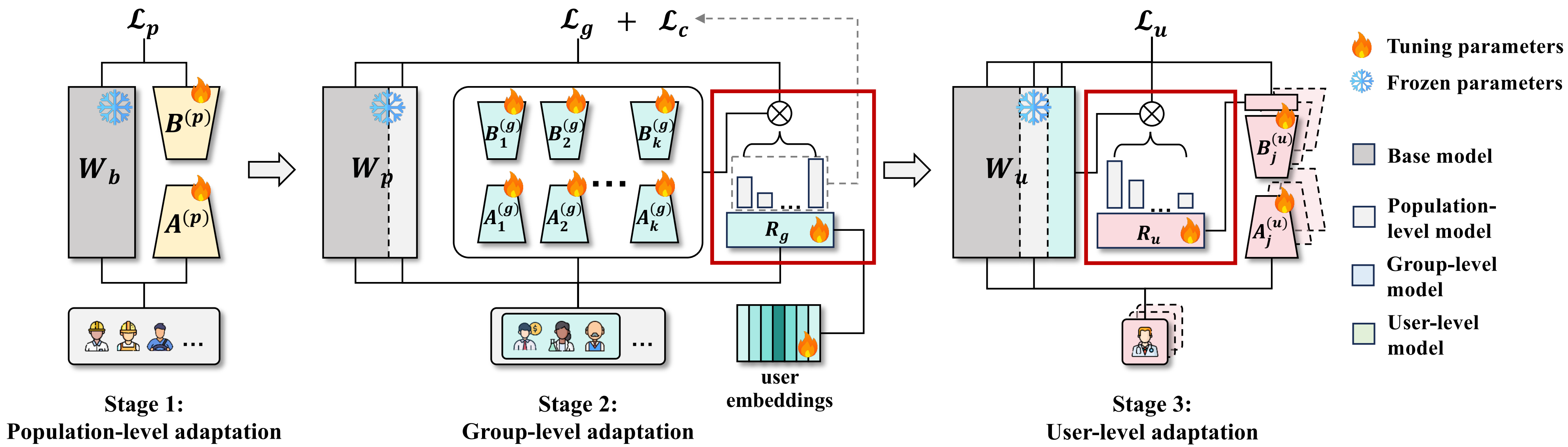


Fig 3. Overview of the training process of PROPER.

Method

Stage 1 (Population-Level Adaptation):

- The update process of the feed-forward network (FFN) block in a Transformer can be expressed as:

$$o = Wx = W_b + \Delta W x$$

- In the population-level adaptation stage, parameter updates are formulated as:

$$o = W_b x + \frac{\alpha}{r} B^{(p)} A^{(p)} x$$

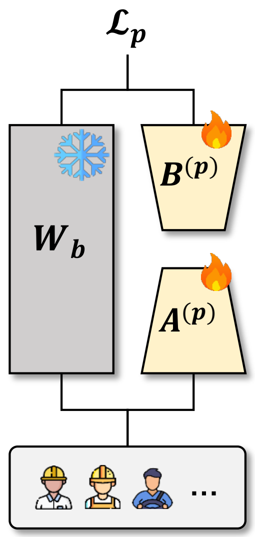
- The population-level LoRA is trained via fine-tuning using the cross-entropy loss:

$$\mathcal{L}_p = \sum_i \mathrm{CE}\{\mathrm{LLM}(q_i|\Omega_p), r_i\}$$

- With the loss, Stage 1 parameters are learned and merged into the backbone parameters for the next training stage:

$$\hat{B}^{(p)}, \hat{A}^{(p)} = \arg\min_{\Omega_p} \mathcal{L}_p$$

$$W_p = W_b + \hat{B}^{(p)} \hat{A}^{(p)}$$



Stage 1:
Population-level adaptation

Method

Stage 2 (Group-Level Adaptation):

- Employing LoRAMoE, represent each group with a LoRA experts:

$$o = W_p x + \sum_{i=1}^{k} \omega_i B_i^{(g)} A_i^{(g)} x$$

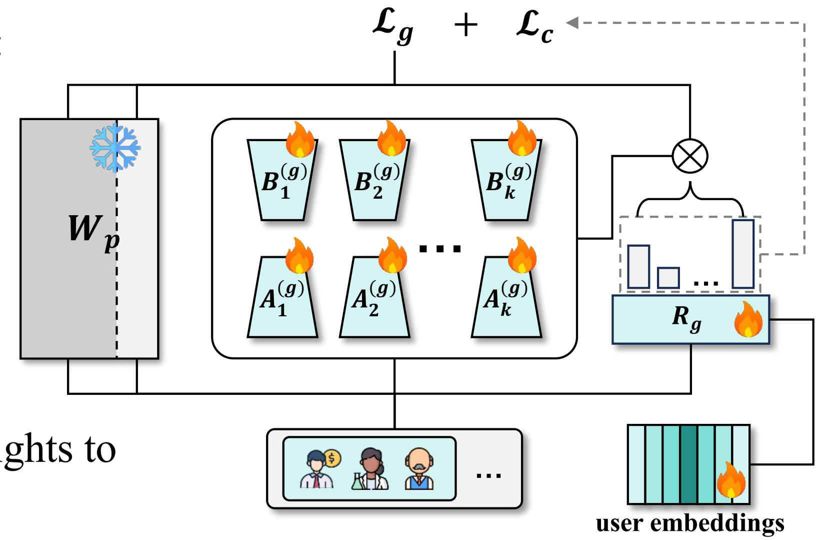
- Assign users to groups dynamically through a user-aware router:

$$\omega(x) = \text{softmax}(h),$$
$$h = \text{softmax}(xM_g) + \text{softmax}(uM_u)$$

- Constraint loss to encourage the router to assign distinct expert weights to different users:

$$s_{(i,j)} = \omega_{u_i}^T \omega_{u_j} \quad \mathcal{L}_c = \sum_{i \neq j} |s_{(i,j)}|$$

- Learn Stage 2 parameters and merge into the backbone parameters with similar process in Stage 1:



Stage 2:
Group-level adaptation

$$\mathcal{L}_g = \sum_i \text{CE}\{\text{LLM}(q_i|\Omega_g), r_i\} \quad \hat{B}_j^{(g)}, \hat{A}_j^{(g)} = \arg\min_{\Omega_g} \mathcal{L}_g \quad W_g = W_p + \frac{a}{r} \sum_{j=1}^{k} \omega_j B_j^{(g)} A_j^{(g)}$$

Method

Stage 3 (User-Level Adaptation):

- Assign a unique LoRA to each user:

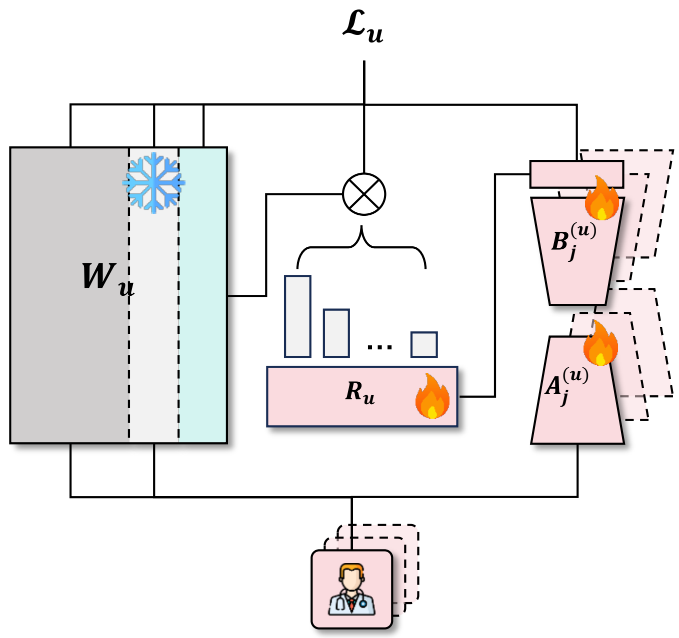
$$o = W_g x + B_j^{(u)} A_j^{(u)} x$$

- A new LoRA-aware router that dynamically integrates group-level LoRAs and user-level LoRAs:

$$\beta_u(x) = \text{softmax}(W_l h_u)$$
$$h_u = \text{LoRA}_u(x),$$

- Learn Stage 3 parameters and merge into the backbone parameters with similar process in Stage 1&2:

$$\mathcal{L}_p^{(j)} = \sum_i \text{CE}\{\text{LLM}(q_i^{(j)} | \Omega_p^{(j)}), r_i^{(j)}\},$$
$$\hat{B}_j^{(u)}, \hat{A}_j^{(u)} = \arg\min_{\Omega_p^{(j)}} \mathcal{L}_p^{(j)},$$
$$W_u^{(j)} = W_g + B_j^{(u)} A_j^{(u)} + \sum_{m=1}^{k} \beta_m B_m^{(g)} A_m^{(g)},$$



Stage 3:
User-level adaptation

Experiments

Evaluation Dataset:

- LaMP benchmark

Baselines:

- Prompt-based
  - In-Context-Learning (ICL)
  - Retrieval-Augmented Generation (RAG)
  - Profile-Augmented Generation (PAG)
- Fine-tuning-based
  - OPPU (kv, mlp)
  - PROPER (Stage 1, 2, 3)

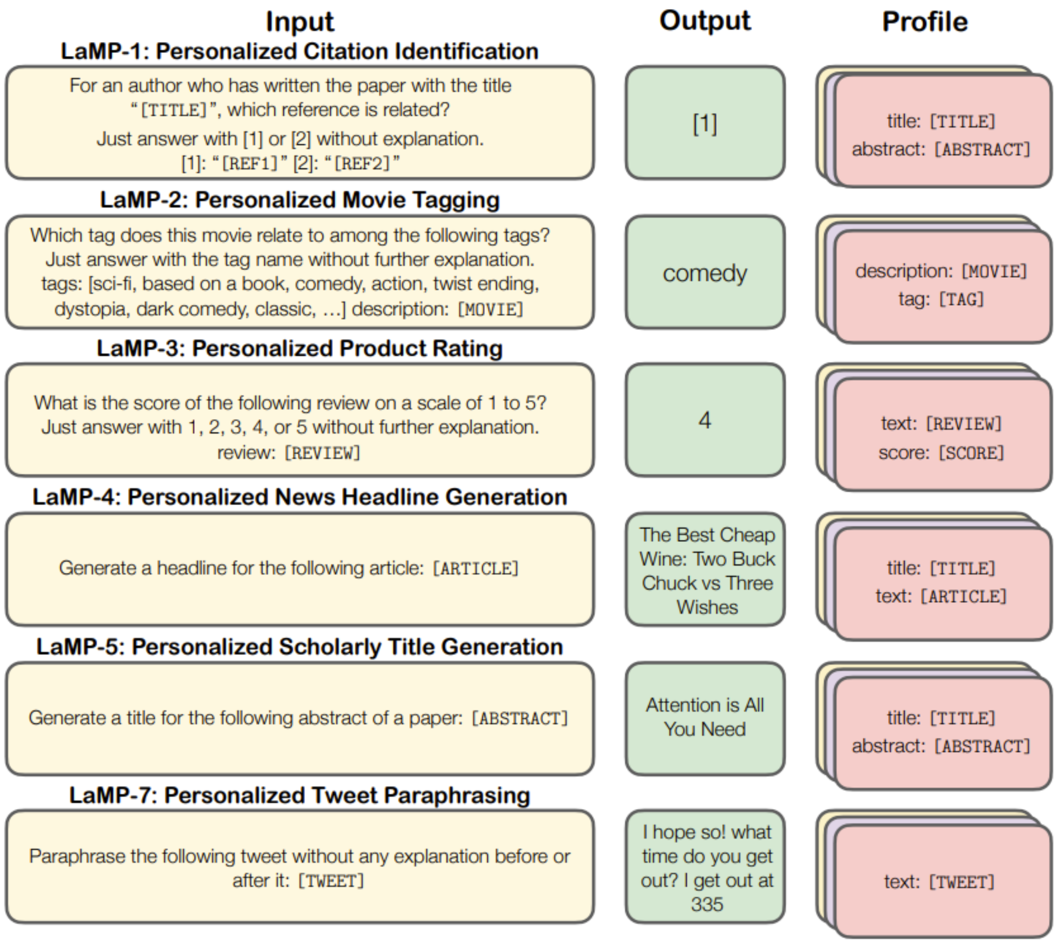


Fig 5. LaMP input & output examples.

Inference Stage

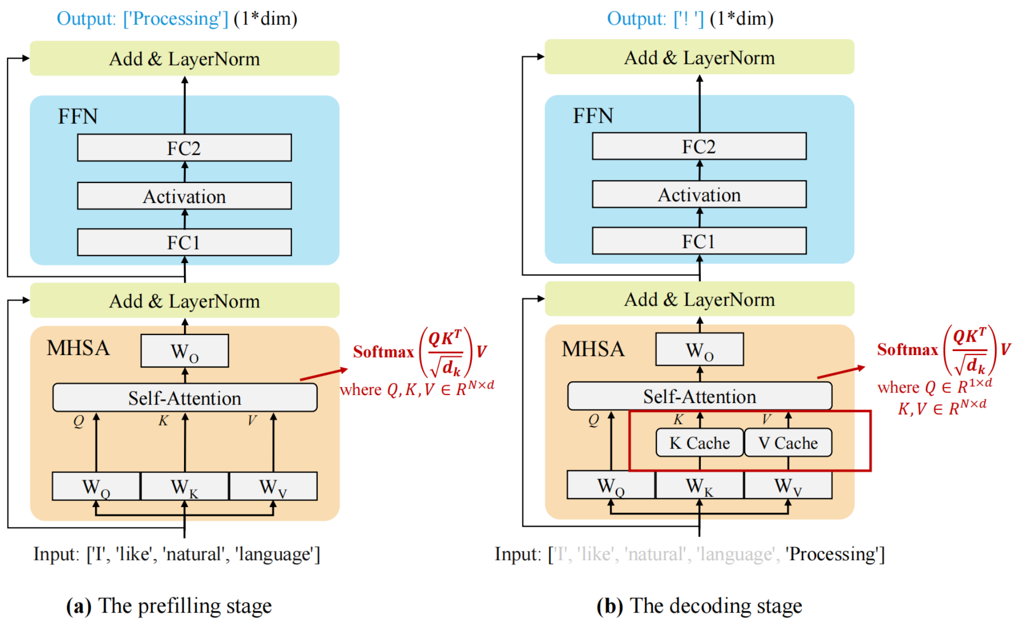


Figure 6: **Autoregressive** token-by-token generation leveraging the **key-value (KV) cache**.

The token-by-token inference of large models results in slower processing speeds, as each step of generation requires the use of the KV-cache.

Multi-output: ToT require independently traversing each branch. This leads to an exponential increase in inference time due to the autoregressive nature of the model.

Long-output: Long-form text leverage the KV-cache to accelerate attention computations, results in a sharp increase in memory usage.

Inference Stage

Multi output

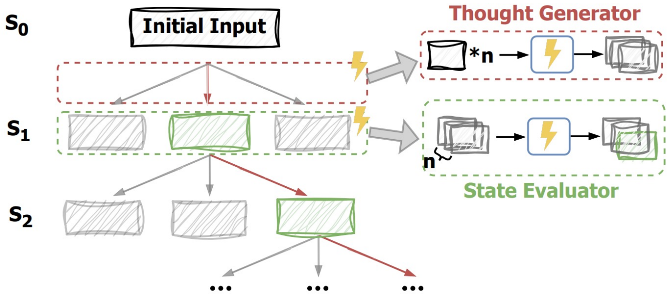


Figure 7: **Tree of thoughts**.

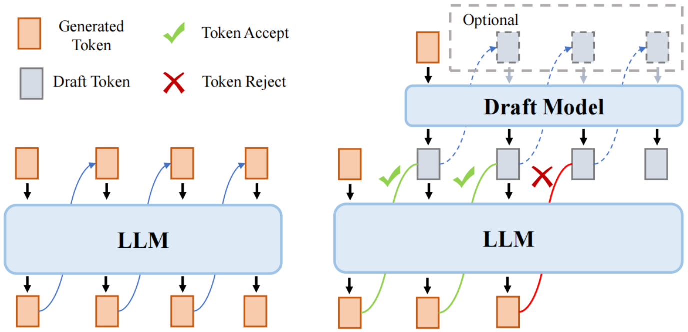


Figure 8: Paradigm of **speculative decoding**.

Long output

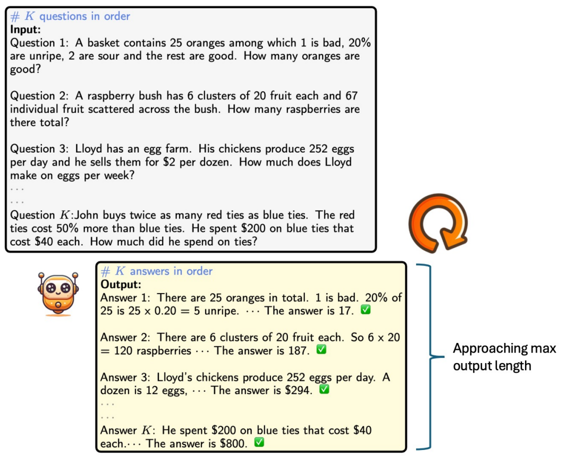


Figure 9: Long-context Generation.

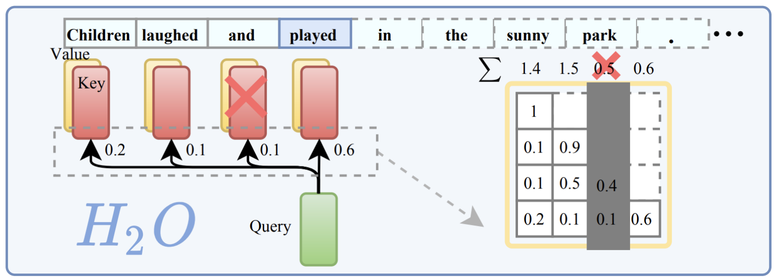


Figure 10: KV cache compression (eviction) guided by attention scores.

Inference Stage

SEED: Accelerating Reasoning Tree Construction via Scheduled

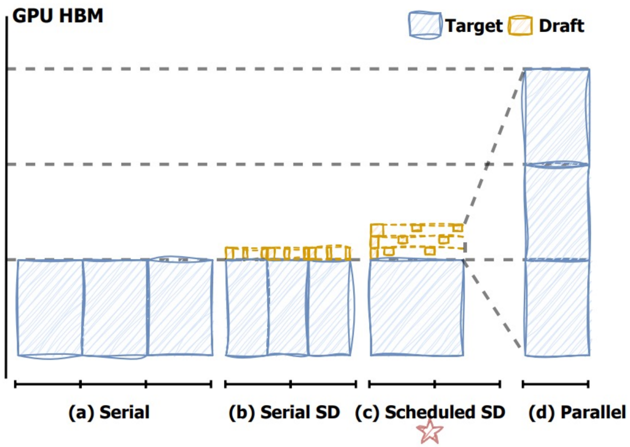


Figure 11: Illustration of four LLM execution strategies for generating 3 sequences in Reasoning Tree construction.

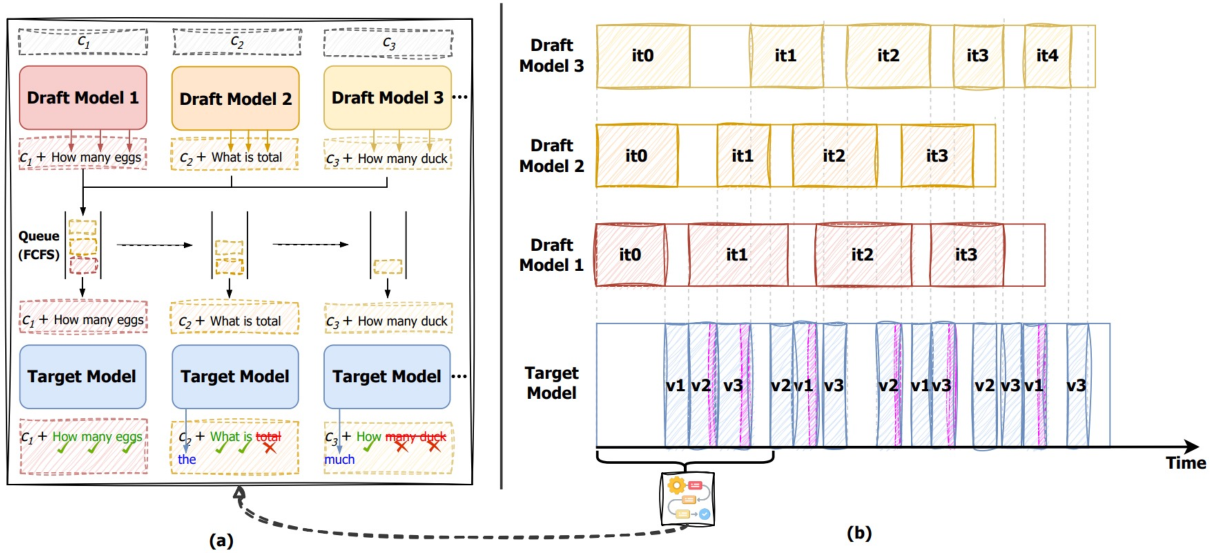


Figure 12: Overall scheduling diagram of **SEED**.

Motivation

- When Large Language Models infer on long-context tasks, the Key-Value (KV) cache occupies ***a larger amount of GPU memory*** and becomes a substantial bottleneck.
- Previous methods fall into two categories:

  (1) The **Prefill-Only** compression method ➡️ memory pressure for long outputs

  (2) The **Unified** compression method ➡️ fine-grained content eviction

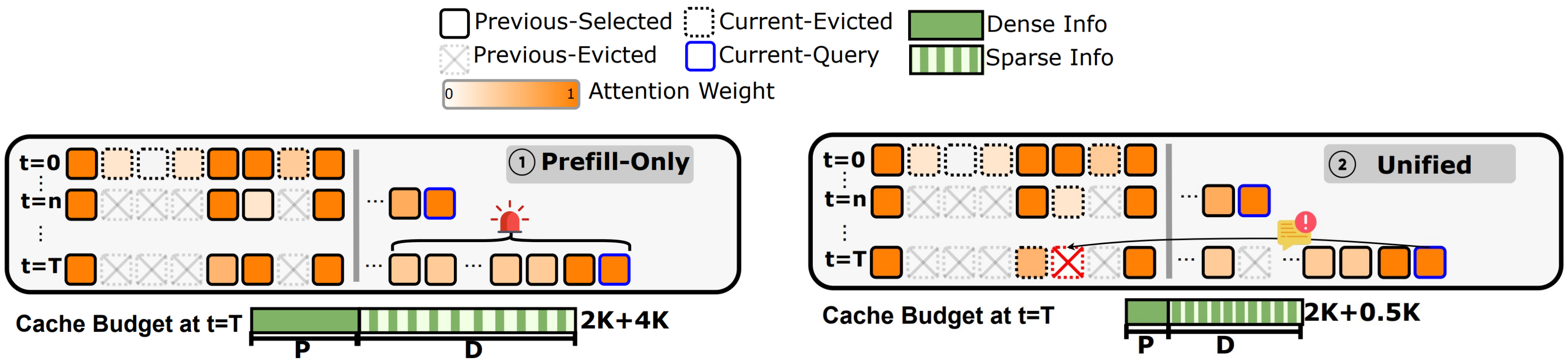


Fig. 1 (a) The Prefill-Only Compression method, *Fig. 1 (b) The Unified Compression method*

Separating the Prefill and Decoding Phases

- **Prefill Phase**: Efficiently preserves essential information in the KV cache during the prefill.

- **Decoding Phase**: Enables optimized allocation of KV cache generated during decoding.

- **Solutions:** SCOPE, a simple yet efficient framework that **S**eparately performs KV **C**ache **O**ptimization during the **P**refill and d**E**coding phases.

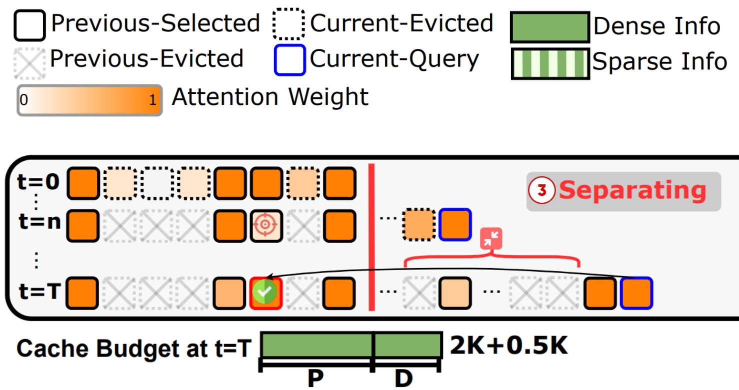


Fig. 1 (c) Separating the prefill and decoding phases

Pilot Observation: KV Cache in Inference Perspective

Prefill Stage

The **20%** compression rate during the prefill phase resulted in nearly **95% degradation** in accuracy on the GSM8k+ task within LONGGENBENCH.

Observations (i): For tasks that require specific fine-grained context, such as reasoning tasks, excessive compression during the prefill phase significantly compromises performance.

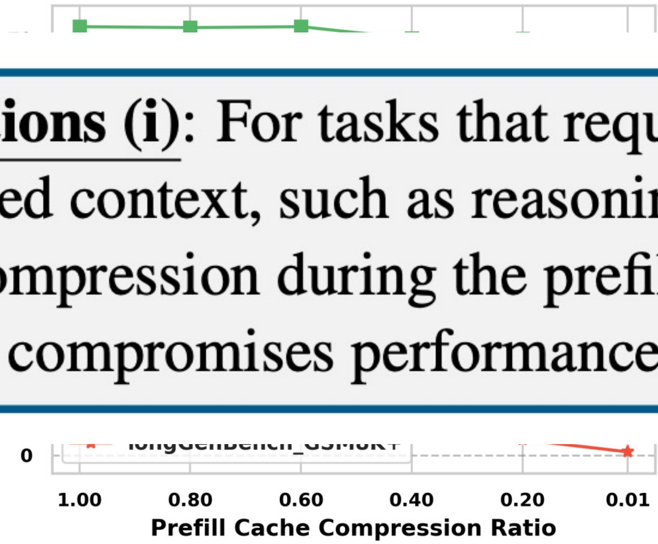


Fig. 2 (a) Performances across various compression ratios during the prefill phase on three tasks under the full decoding cache condition.

Pilot Observation: KV Cache in Inference Perspective

Decoding Stage

Across all three layers, the retained heavy hitters predominantly originate from the KV cache generated **during the decoding phase**.

Observations (ii): During the decoding phase of long text generation, the use of the greedy algorithm may lead to a deviation in heavy hitters.

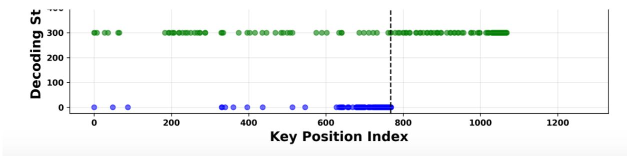


Fig. 2 (a) Position distribution of the heavy hitters, selected by top 15% attention scores, at decoding steps 1, 300, and 500 across layers 0, 13, and 31.

Pilot Observation: KV Cache Budget Reallocation

Observations (i): For tasks that require specific fine-grained context, such as reasoning tasks, excessive compression during the prefill phase significantly compromises performance.

Observations (ii): During the decoding phase of long text generation, the use of the greedy algorithm may lead to a deviation in heavy hitters.

Pilot Observation: KV Cache Budget Reallocation

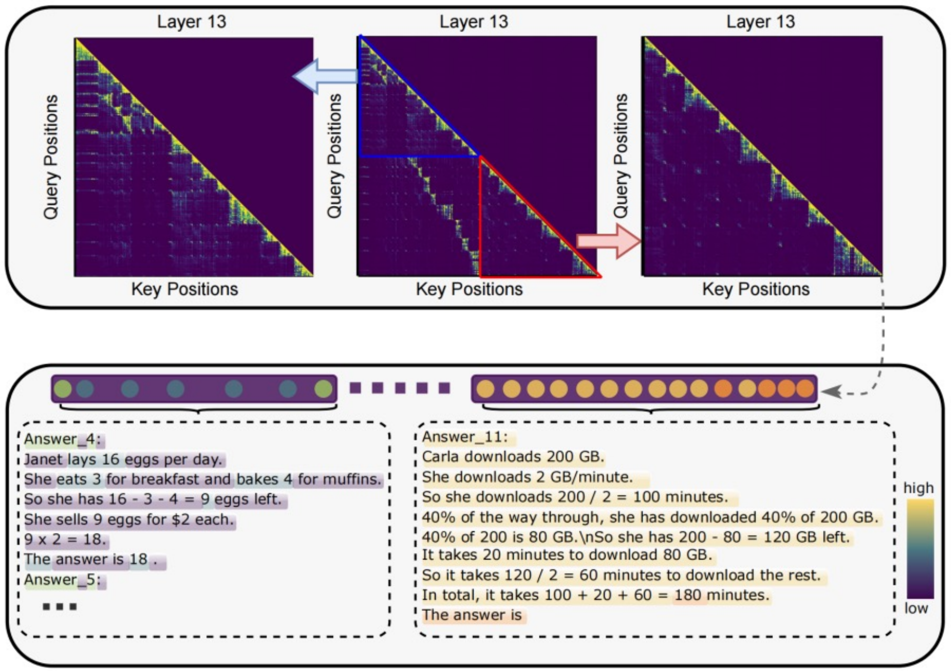
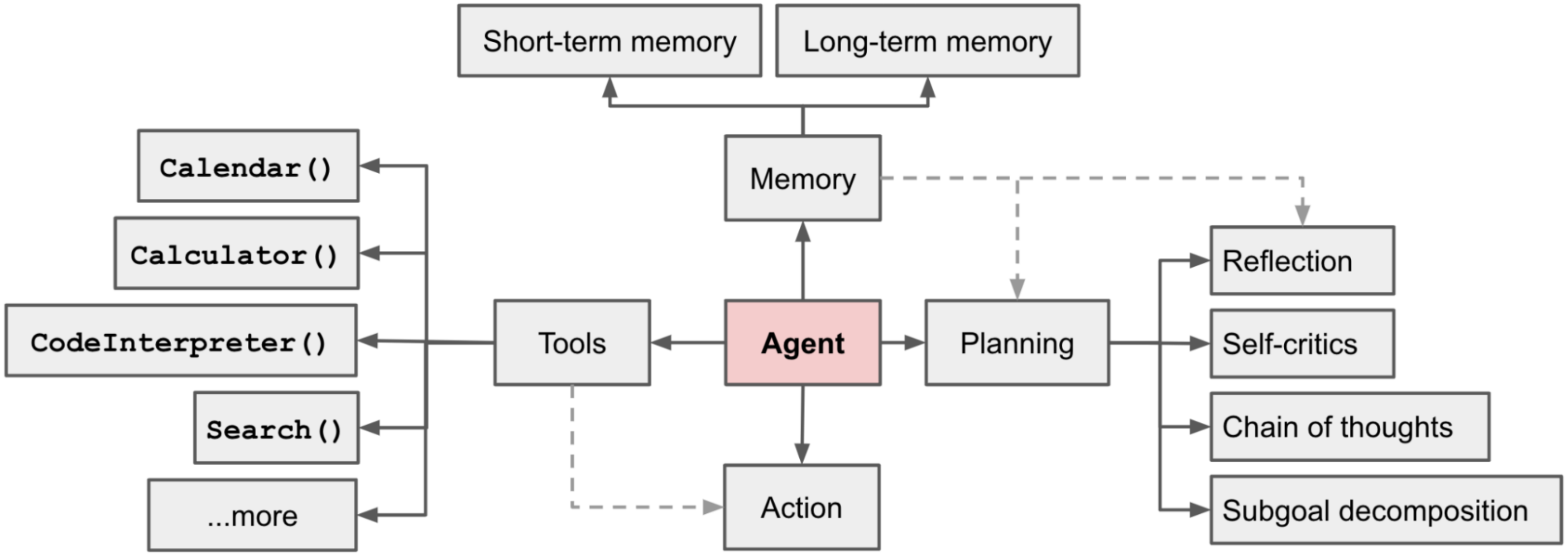


Fig. 2 (c) Attention heatmaps for layer 13 of a GSM8k+ sample in LONGGENBENCH and details of the correspondence between attention scores and generated token positions.

Agent


Short-term memory
Long-term memory
Memory
Calendar()
Calculator()
CodeInterpreter()
Search()
...more
Tools
Agent
Planning
Reflection
Self-critics
Chain of thoughts
Subgoal decomposition
Action

Agent

Agents: An Open-source Framework for Autonomous Language Agents

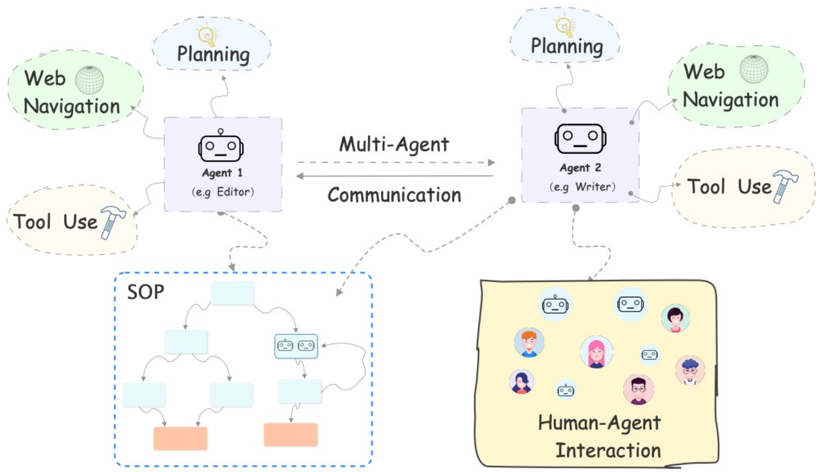


Figure 16: Components in agents framework.

- **Long-short Term Memory**: Long-term memory implemented via VectorDB + Semantic Search and short-term memory (working memory) maintained and updated by an LLM.
- **Tool Usage**: Use any external tools via function-calling.
- **Web Navigation**: Use search engines to navigate the web and get useful information.
- **Multi-agent Communication**
- **Human-Agent interaction**
- **Symbolic Control**: **SOP (Standard Operation Process)** that defines subgoals/subtasks for the overall task to customize fine-grained workflows for the language agents.

Agent

Brief History of Web Agents
Book the cheapest one-way flight from: NLG to: Brownsville, TX on 12/10/2016.
Book Your One-Way Flight
From:
To:
Departure Date
Search
MiniWoB(++)
2017-2018
07/2022
WebShop
[richer env]
Mind2Web
[LLM, generalist, real websites]
06/2023
07/2023
Carnegie Mellon University
WebArena
[sandbox w/ real websites]
Self-hosted fully functional web applications
OneStopShop
CMS
reddit
GitLab
Toolbox
Knowledge resources
WebArena
LMM
Grounding
SeeAct
SeeAct
[multimodal LLM, online eval]
01/2024
VisualWebArena
SeeClick
WebVoyger
WebLINX
WorkArena
Mind2Web-Live
...

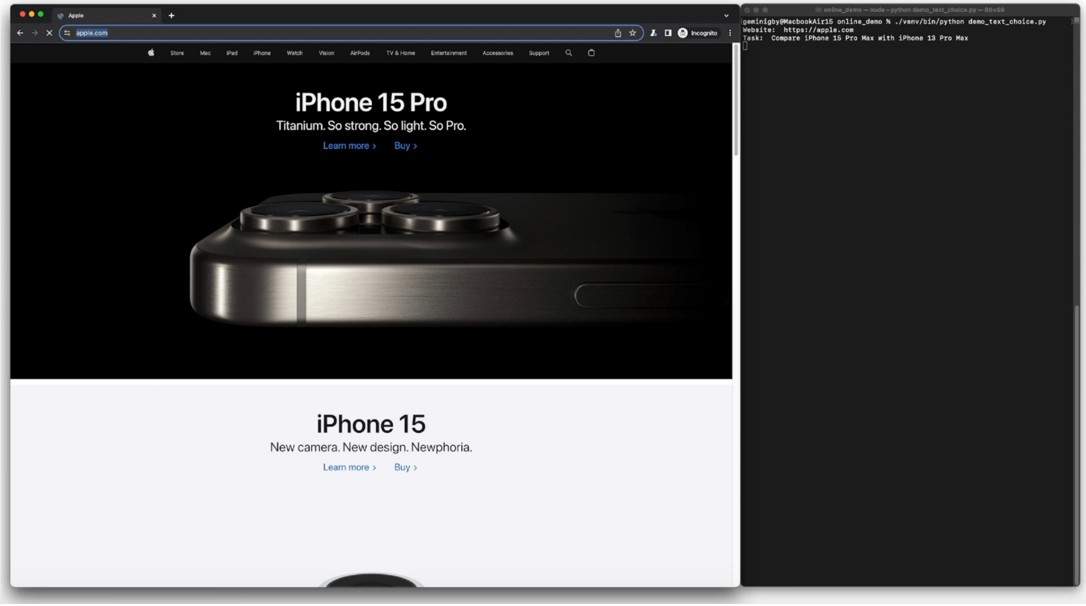
Agent


Brief History of Web Agents
Embodiment: Increasingly human-like and vision-centric
MiniWoB(++)
2017-2018
07/2022
WebShop
[richer env]
Mind2Web
[LLM, generalist, real websites]
06/2023
07/2023
Carnegie Mellon University
WebArena
[sandbox w/ real websites]
SeeAct
[multimodal LLM, online eval]
01/2024
VisualWebArena
SeeClick
WebVoyger
WebLINX
WorkArena
Mind2Web-Live
...
See Also:
Pix2Act
AgentBench
CogAgent
RCI-Agent
Dual-VCR
WebAgent
Book Your One-Way Flight
From:
To:
Departure Date
Search
Self-hosted fully functional web applications
OneStopShop
CMS
reddit
GitLab
Toolbox
Knowledge resources
Environment: Increasingly broad, complex, and realistic

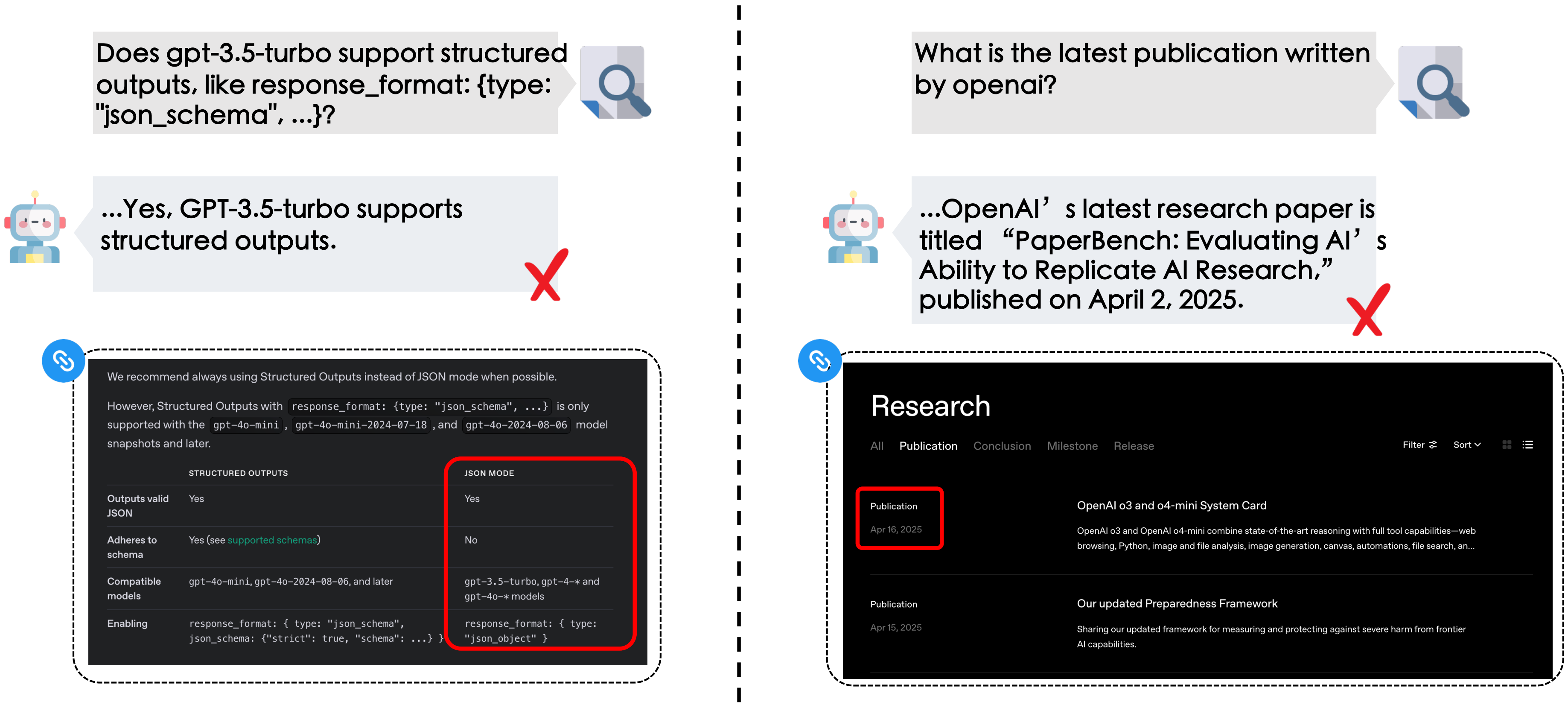
Agent

(NeurIPS'23) (ICML'24)

Generalist Web Agents: Mind2Web & SeeAct



Website: https://apple.com
Task: Compare iPhone 15 Pro Max with iPhone 13 Pro Max



- Can not find **information hidden within deep webpages** or **the latest updates**.

Agent

How can an agent navigate across web pages to seek deep, non-obvious information?

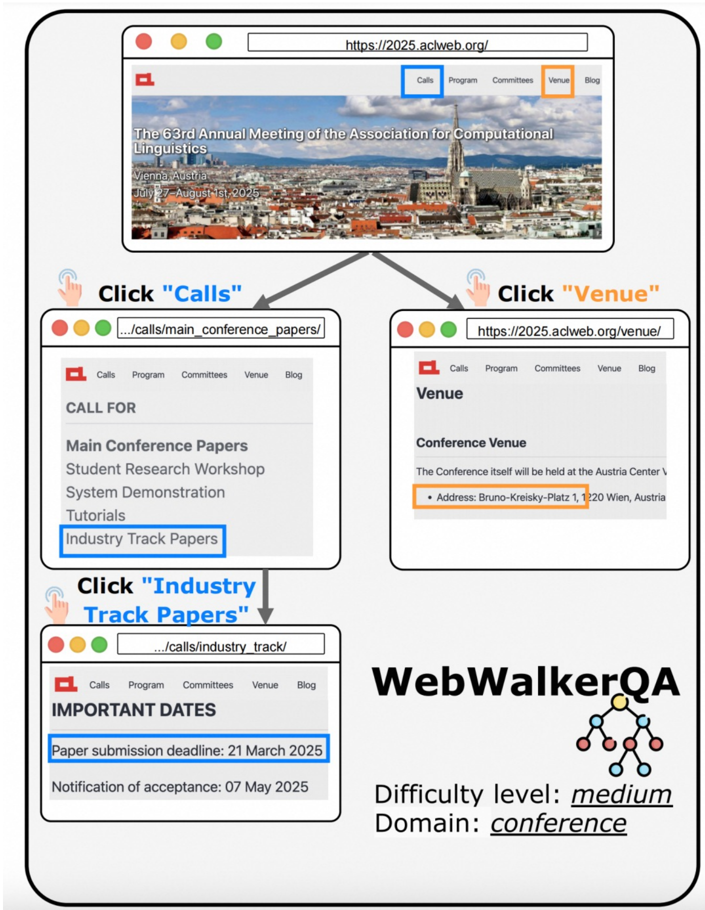
- Unlock the capability of persistent, multi-hop web exploration

Motivation

Key challenge in RAG:
Traditional online search may not trace the **Deeper content** embedded within website.

When is the paper **submission deadline for the ACL 2025 Industry Track**, and what is the **venue address for the conference**?

https://2025.aclweb.org/



Motivation

How to solve it:

Interacting with the web pages and **digging through** them can effectively address **deep information seeking**.

We constrain actions to cli to evaluate the agent's navigation and information-seeking capabilities.

- We propose **Web Traversal task**.
- We construct a challenging benchmark, **WebWalkerQA**.
- To tackle the challenge of web-navigation tasks requiring long context, we propose **WebWalker**.

Datasets

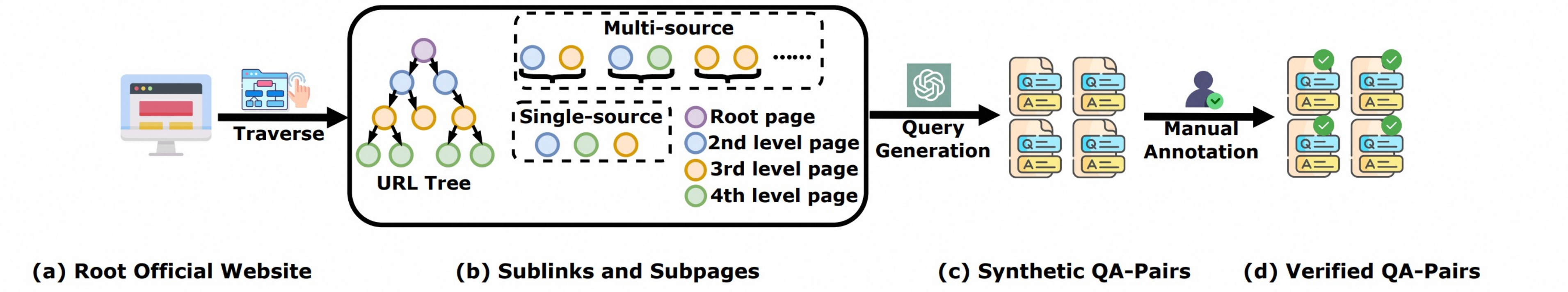
WebWalkerQA

	Language	Format	Depth	Width	Hop	# Pages
Mind2Web (Deng et al., 2023)	En	Multi-choice	✗	✗	✗	100
WebArena (Zhou et al., 2024a)	En	Action	✗	✗	✗	6
AssistantBench (Yoran et al., 2024)	En	QA	✗	✓	✓	525
MMInA (Zhang et al., 2024c)	En	Action	✗	✓	✓	100
GAIA (Mialon et al., 2024)	En	QA	✗	✓	✓	-
WebWalkerQA	En&Zh	QA	✓	✓	✓	1373

Comparison between WebWalkerQA and other benchmarks.

Datasets

WebWalkerQA



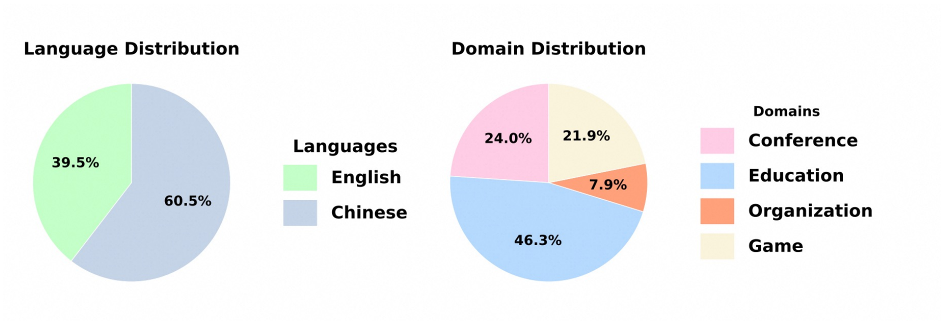
Data Generation Pipeline for WebWalkerQA.

Datasets

WebWalkerQA

Single-source QAs			Multi-source QAs		
Easy	Medium	Hard	Easy	Medium	Hard
80	140	120	80	140	120

Dataset statistics on difficulty level.



Language and domain **distribution**.

Datasets

Web Traversal Task:

Given an initial website URL and a query Q, which needs to be answered by exploring the website. The goal of this task is to gather enough information through page traversal to ultimately answer the query Q.

Evaluation:

Correctness -> *acc.* Evaluated by GPT-4o
Efficiency -> Action count of successful agentic executions

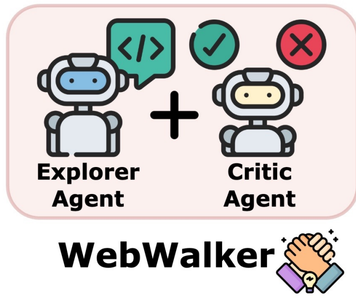
When is the paper **submission deadline for the ACL 2025 Industry Track**, and what is the **venue address for the conference**?

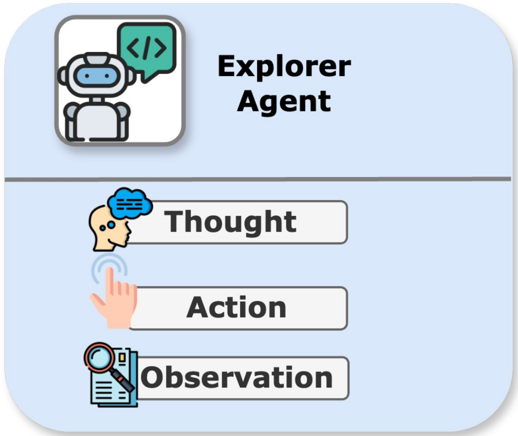
https://2025.aclweb.org/

The paper submission deadline for the ACL 2025 Industry Track is **March 21, 2025** and the conference will be held in **Bruno-Kreisky-Platz 1**.

Methods

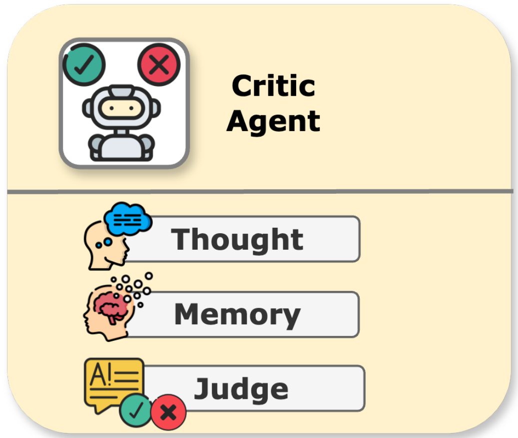
WebWalker: a multi-agent framework





Think then Explore

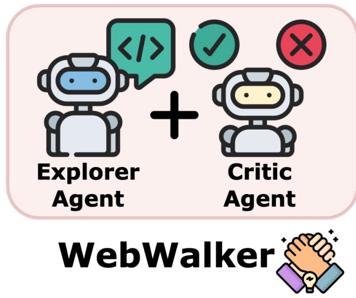
ReAct format

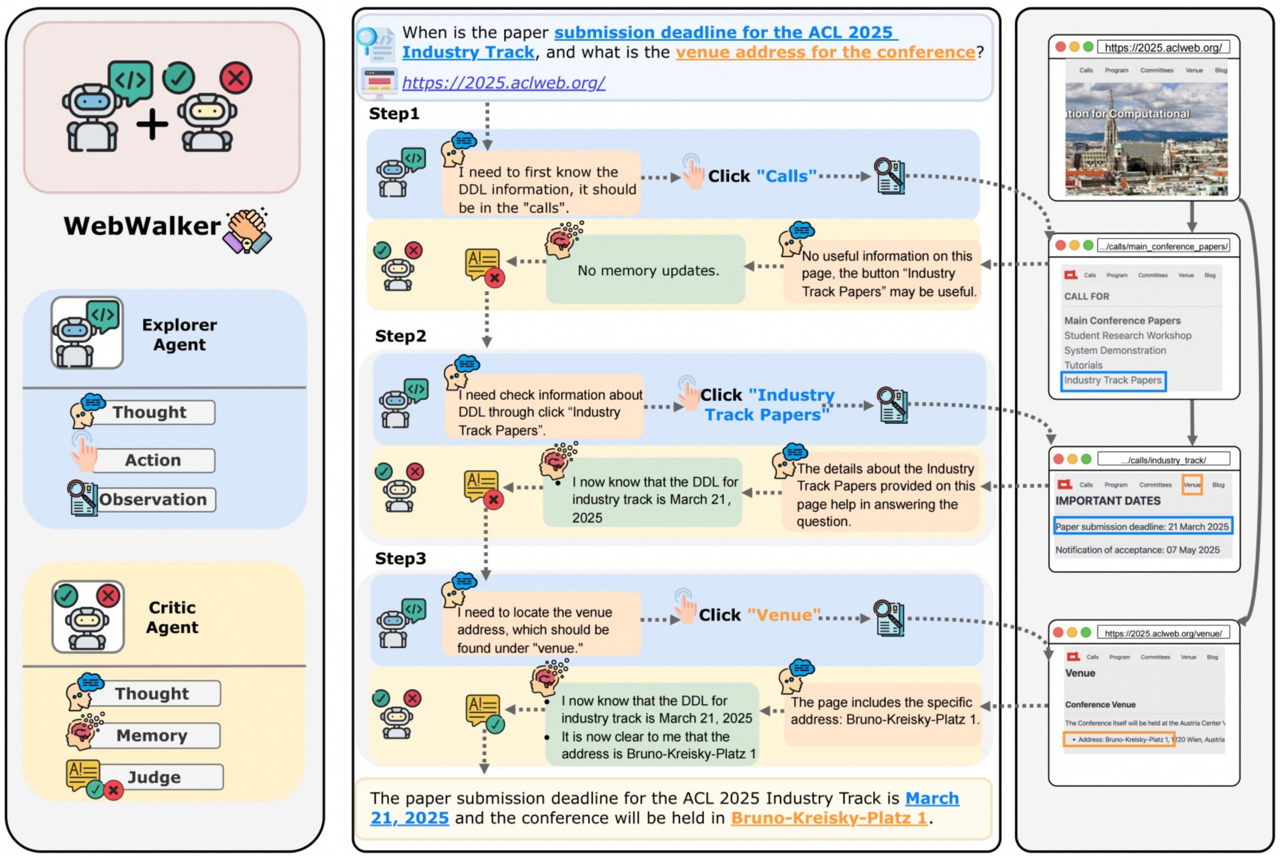


Think then Critique

Motivated by pair programming

Methods



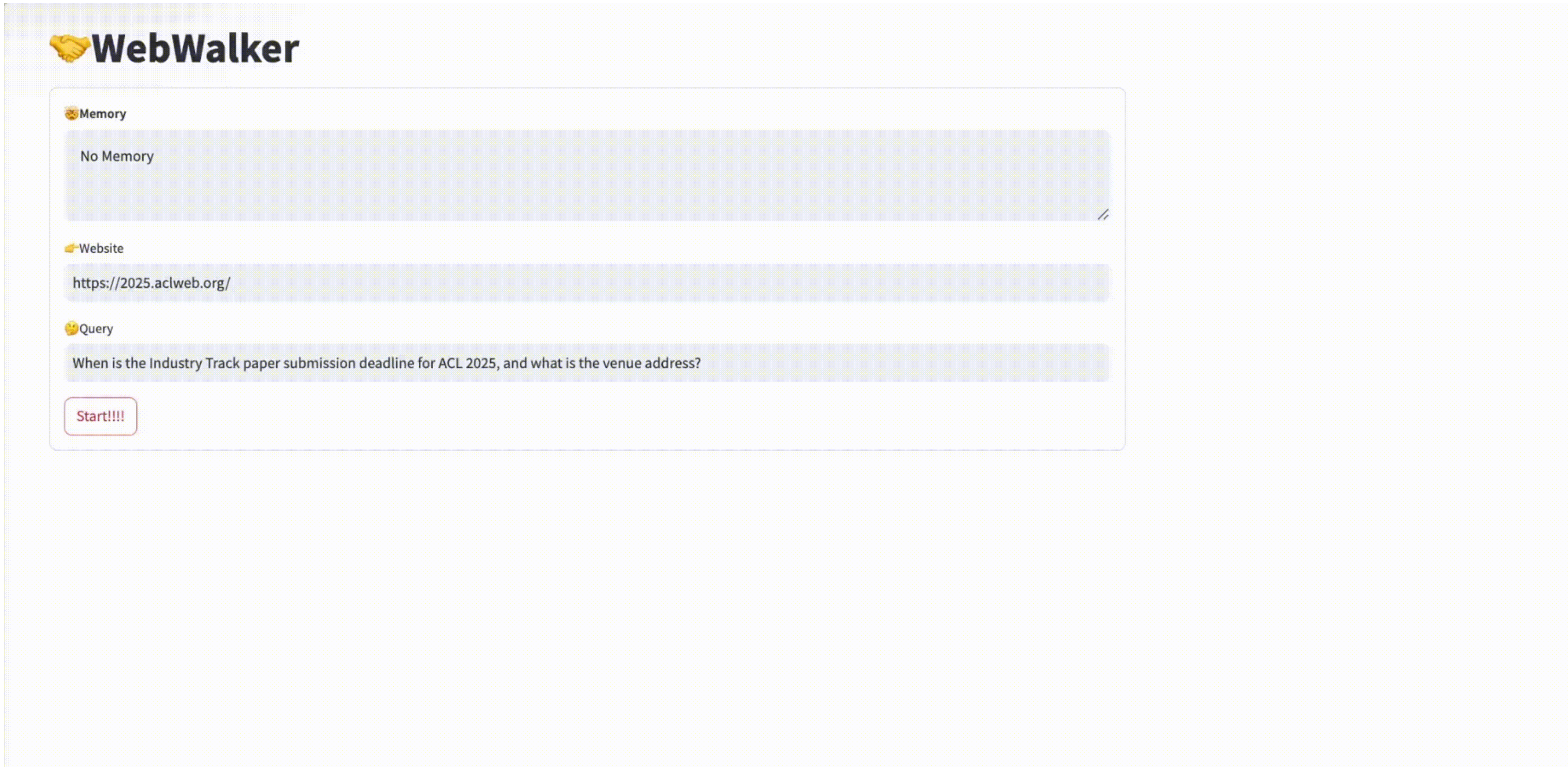


The explorer agent traverses the web pages in **Thought-Action-Observation** (T, A, O) paradigms.

The critic agent **updates the memory** until sufficient information is accumulated to effectively **address the query**.

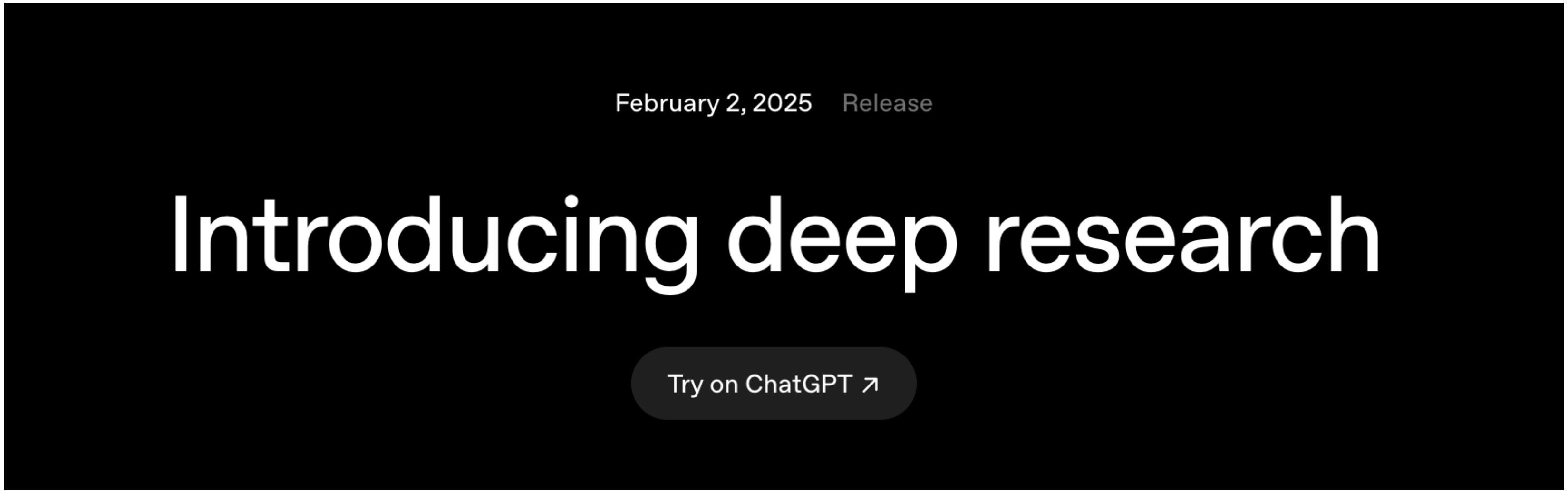
Methods

Explorer Agent
Critic Agent
WebWalker

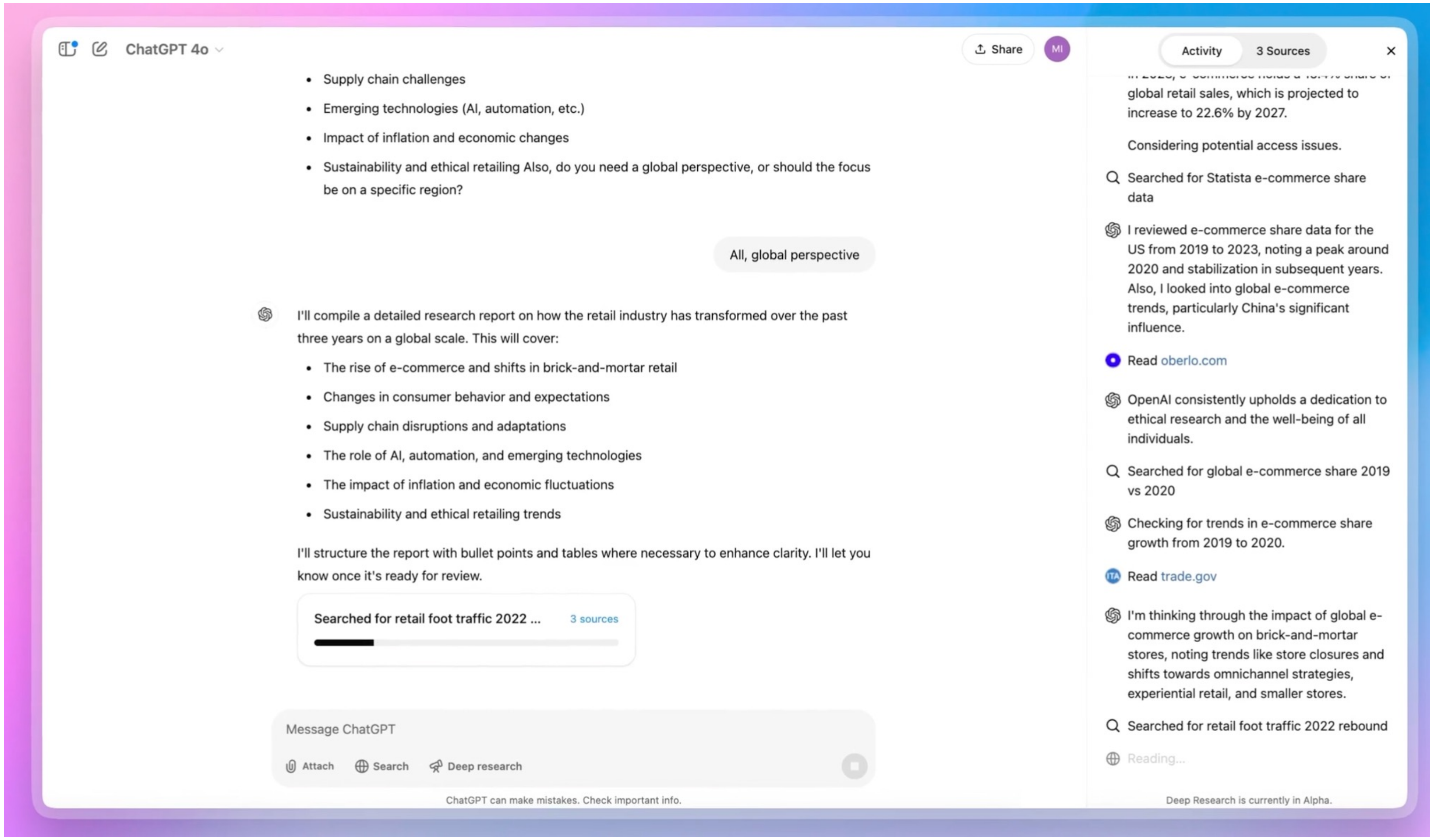


WebWalker
Memory
No Memory
Website
https://2025.aclweb.org/
Query
When is the Industry Track paper submission deadline for ACL 2025, and what is the venue address?
Start!!!!

Agent


February 2, 2025
Release
Introducing deep research
Try on ChatGPT ↗

Agent


ChatGPT 4o
Share
Supply chain challenges
Emerging technologies (AI, automation, etc.)
Impact of inflation and economic changes
Sustainability and ethical retailing Also, do you need a global perspective, or should the focus be on a specific region?
All, global perspective
I'll compile a detailed research report on how the retail industry has transformed over the past three years on a global scale. This will cover:
The rise of e-commerce and shifts in brick-and-mortar retail
Changes in consumer behavior and expectations
Supply chain disruptions and adaptations
The role of AI, automation, and emerging technologies
The impact of inflation and economic fluctuations
Sustainability and ethical retailing trends
I'll structure the report with bullet points and tables where necessary to enhance clarity. I'll let you know once it's ready for review.
Searched for retail foot traffic 2022 ...
3 sources
Message ChatGPT
Attach
Search
Deep research
ChatGPT can make mistakes. Check important info.
Activity
3 Sources
global retail sales, which is projected to increase to 22.6% by 2027.
Considering potential access issues.
Searched for Statista e-commerce share data
I reviewed e-commerce share data for the US from 2019 to 2023, noting a peak around 2020 and stabilization in subsequent years. Also, I looked into global e-commerce trends, particularly China's significant influence.
Read oberlo.com
OpenAI consistently upholds a dedication to ethical research and the well-being of all individuals.
Searched for global e-commerce share 2019 vs 2020
Checking for trends in e-commerce share growth from 2019 to 2020.
Read trade.gov
I'm thinking through the impact of global e-commerce growth on brick-and-mortar stores, noting trends like store closures and shifts towards omnichannel strategies, experiential retail, and smaller stores.
Searched for retail foot traffic 2022 rebound
Reading...
Deep Research is currently in Alpha.

Agent

How to build a web agent like ***Deep Research*** from scratch?

- unlock the autonomous multi-turn information seeking agency

Pipelines

Step I: Construct diverse and challenging deep information seeking QA pairs based on the real-world web environment;

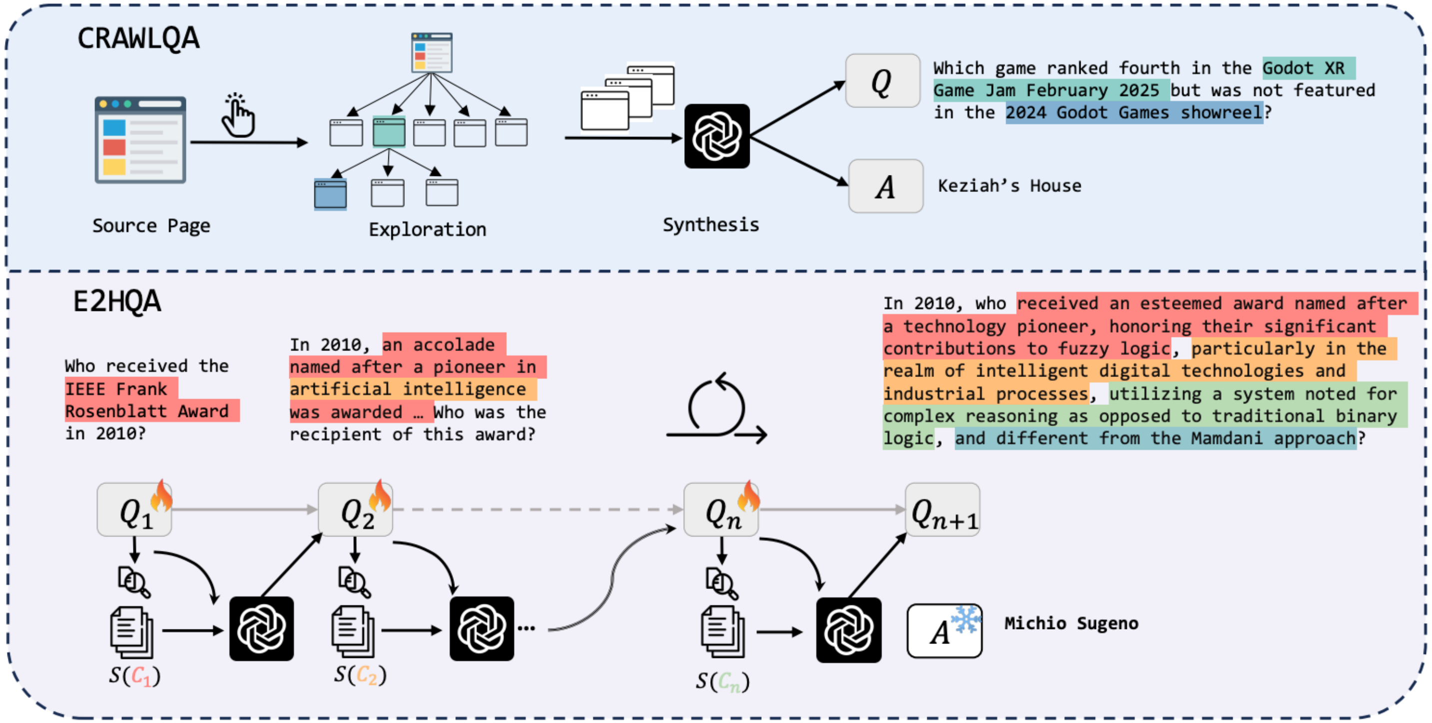
Step II: Sample high-quality trajectories from QA pairs using both LLMs and LRMs to guide the agency learning process;

Step III: Perform fine-tuning to adapt the format instruction following to agentic tasks and environments;

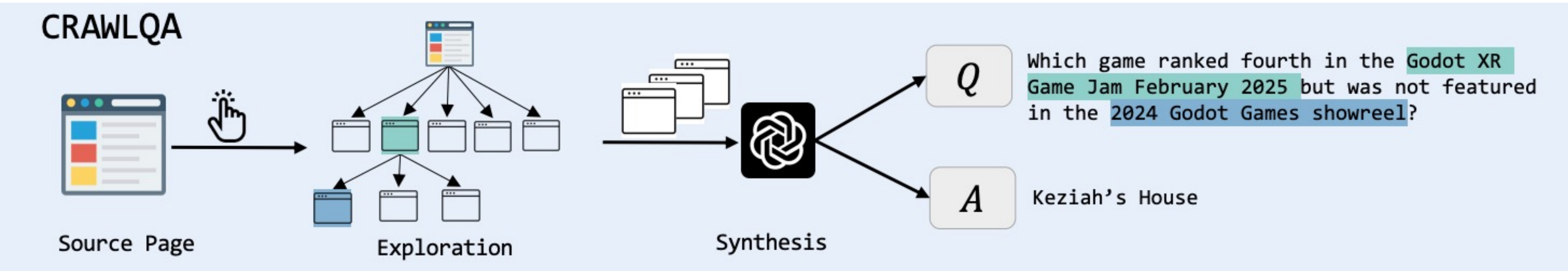
Step IV: Apply RL to optimize the agent's decision-making and generalization capabilities in real-world web environments

Datasets

Previous training datasets are relatively simple and do not capture the real-world challenges.

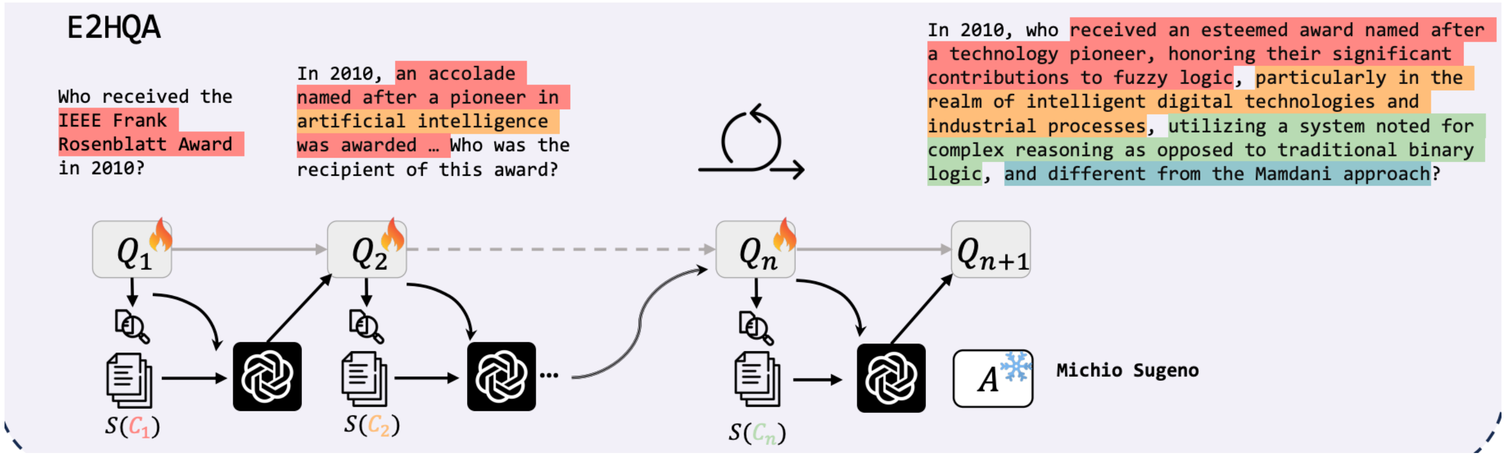


Datasets



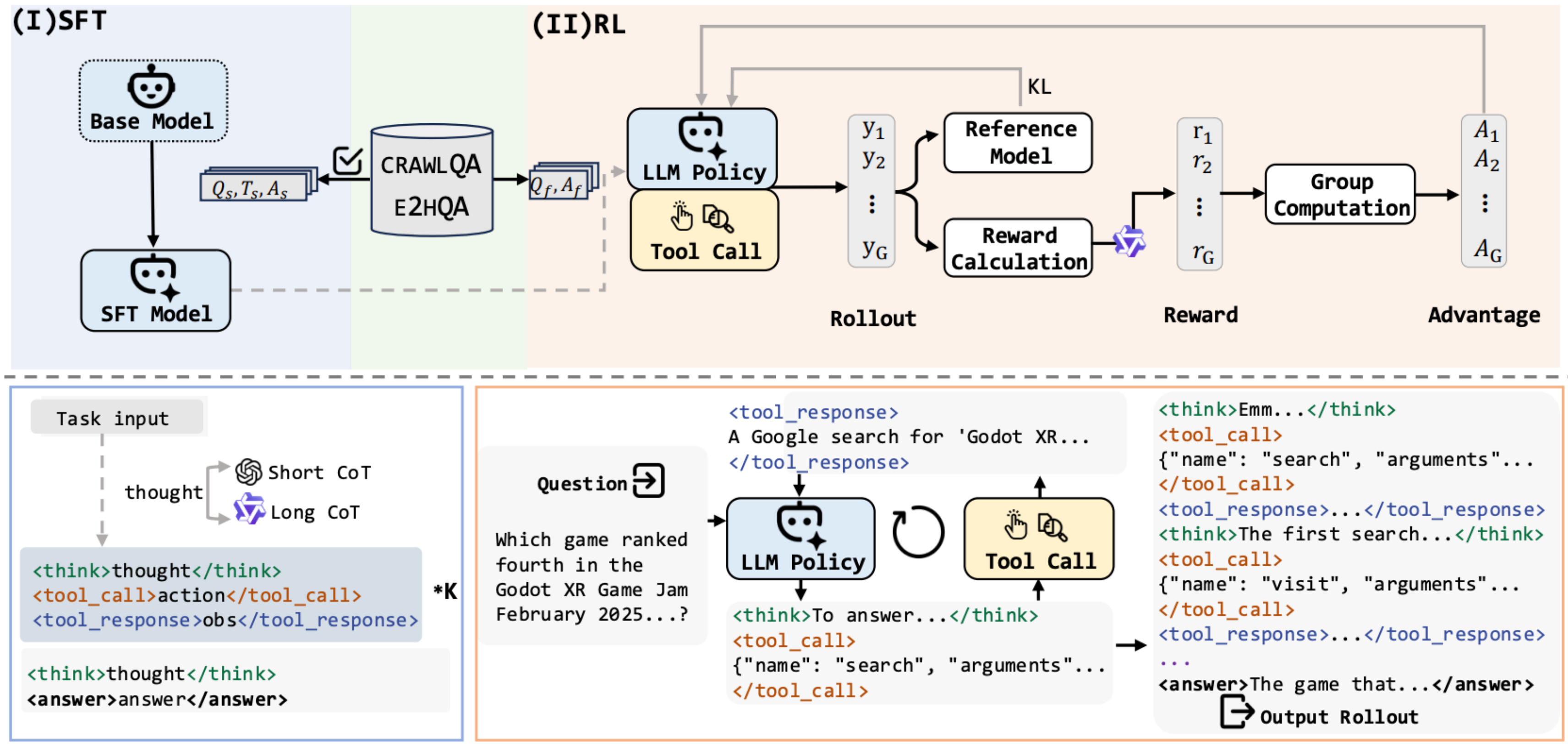
Mimic human behavior by systematically clicking and collecting subpages accessible through sublinks.

Datasets



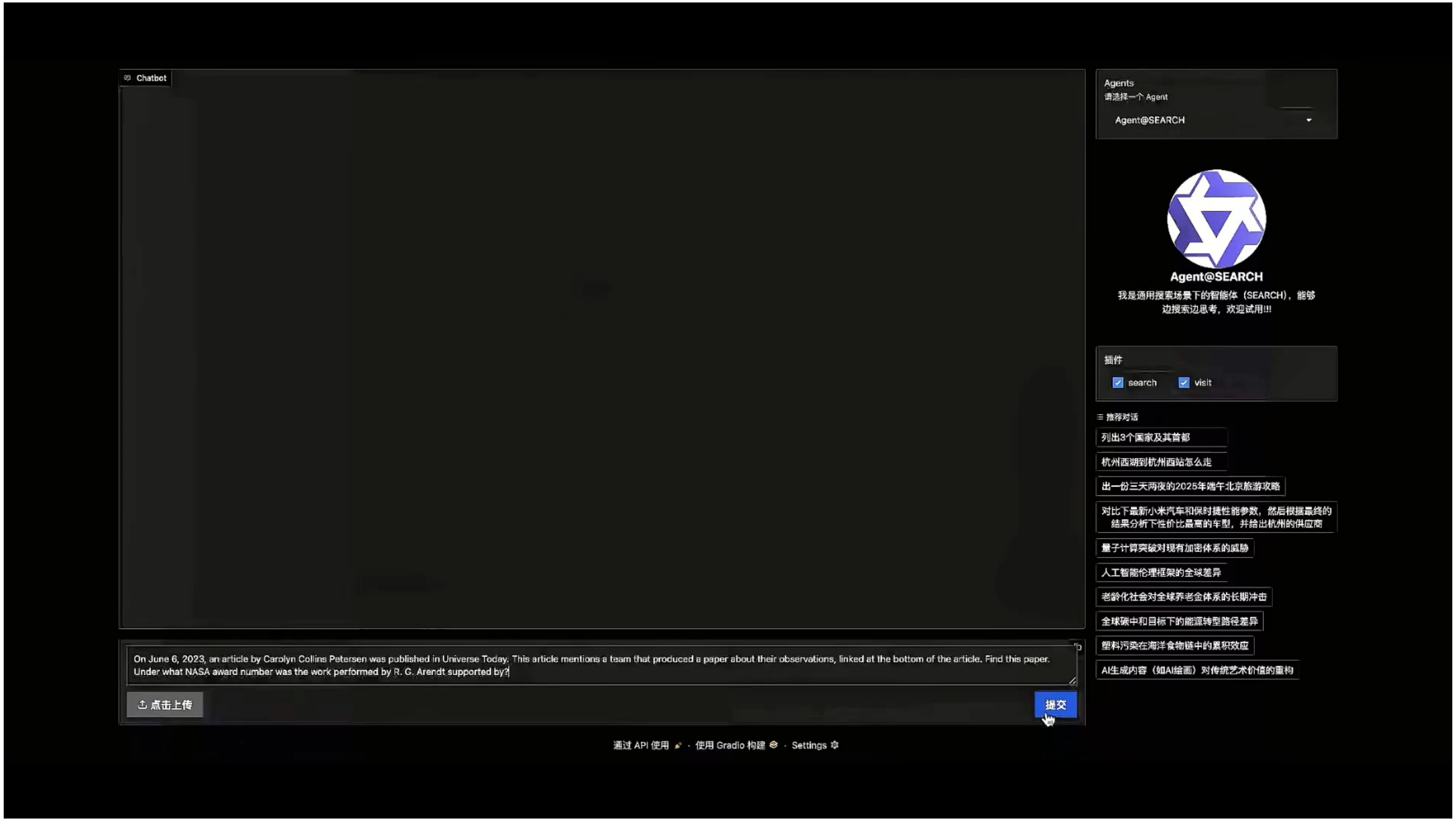
Rewrite simple questions into more complex, challenging ones systematically .

Methods



(I)SFT
Base Model
SFT Model
Q_s, T_s, A_s
CRAWLQA
E2HQA
Q_f, A_f
(II)RL
LLM Policy
Tool Call
KL
y_1 y_2 ⋮ y_G
Reference Model
Reward Calculation
r_1 r_2 ⋮ r_G
Group Computation
A_1 A_2 ⋮ A_G
Rollout
Reward
Advantage
Task input
thought
Short CoT
Long CoT
<think>thought</think>
<tool_call>action</tool_call>
<tool_response>obs</tool_response>
*K
<think>thought</think>
<answer>answer</answer>
Question
Which game ranked fourth in the Godot XR Game Jam February 2025...?
<tool_response>
A Google search for 'Godot XR...
</tool_response>
LLM Policy
Tool Call
<think>To answer...</think>
<tool_call>
{"name": "search", "arguments"...
</tool_call>
<think>Emm...</think>
<tool_call>
{"name": "search", "arguments"...
</tool_call>
<tool_response>...</tool_response>
<think>The first search...</think>
<tool_call>
{"name": "visit", "arguments"...
</tool_call>
<tool_response>...</tool_response>
...
<answer>The game that...</answer>
Output Rollout

Demos


Chatbot
Agents
请选择一个 Agent
Agent@SEARCH
Agent@SEARCH
我是通用搜索场景下的智能体（SEARCH），能够
边搜索边思考，欢迎试用!!!
插件
search
visit
推荐对话
列出3个国家及其首都
杭州西湖到杭州西站怎么走
出一份三天两夜的2025年端午北京旅游攻略
对比下最新小米汽车和保时捷性能参数，然后根据最终的
结果分析下性价比最高的车型，并给出杭州的供应商
量子计算突破对现有加密体系的威胁
人工智能伦理框架的全球差异
老龄化社会对全球养老金体系的长期冲击
全球碳中和目标下的能源转型路径差异
塑料污染在海洋食物链中的累积效应
AI生成内容（如AI绘画）对传统艺术价值的重构
On June 6, 2023, an article by Carolyn Collins Petersen was published in Universe Today. This article mentions a team that produced a paper about their observations, linked at the bottom of the article. Find this paper.
Under what NASA award number was the work performed by R. G. Arendt supported by?
点击上传
提交
通过 API 使用 · 使用 Gradio 构建 · Settings

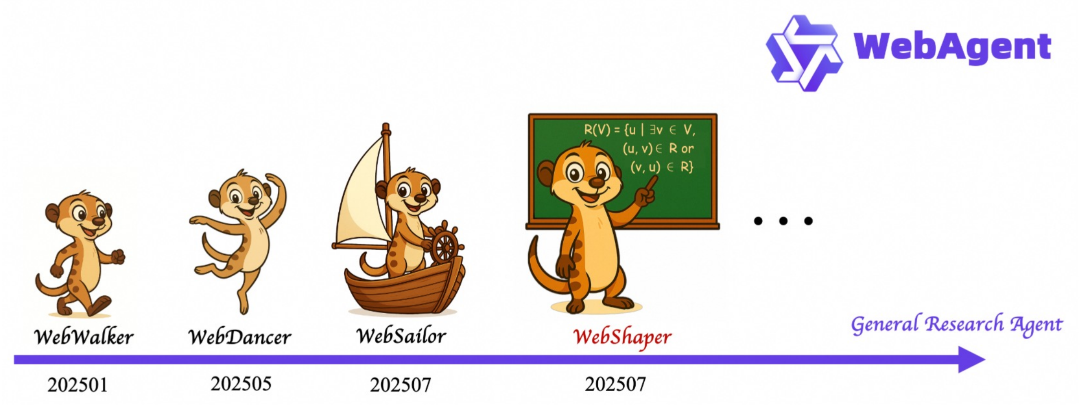
WebAgent

WebAgent for Information Seeking built by Tongyi Lab, Alibaba Group

GITHUB TRENDING
#1 Repository Of The Day

WebSailor-3B | ModelScope WebSailor-3B |

WebDancer-QwQ-32B | ModelScope WebDancer-QwQ-32B | WebWalkerQA



You can check the paper of WebDancer and WebWalker and WebSailor and WebShaper.

GITHUB TRENDING
#1 Repository Of The Day

https://github.com/Alibaba-NLP/WebAgent

If you like our project, feel free to give us a ⭐ on GitHub!

Thanks for watching!
QA