

Introduction

Does gpt-3.5-turbo support structured outputs, like response for mat: {type: "json_schema", ...}?



...Yes, GPT-3.5-turbo supports structured outputs. ✗


What is the latest publication written by openai?



...OpenAI's latest research paper is "PaperBench: Evaluating AI's Ability to Replicate AI Research," published on April 2, 2025.

How to solve it:

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

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

Web Traversal Task

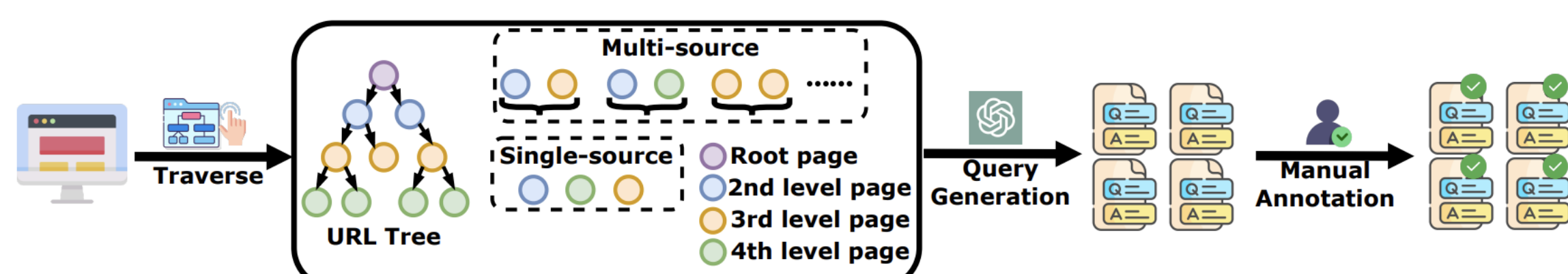
Given a URL root and a query, the goal of this task is to gather enough information through page traversal to ultimately answer the query.

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/>

Traditional online search may **not** trace the **Deeper content** embedded within website.

WebWalkerQA



(a) Root Official Website

(b) Sublinks and Subpages

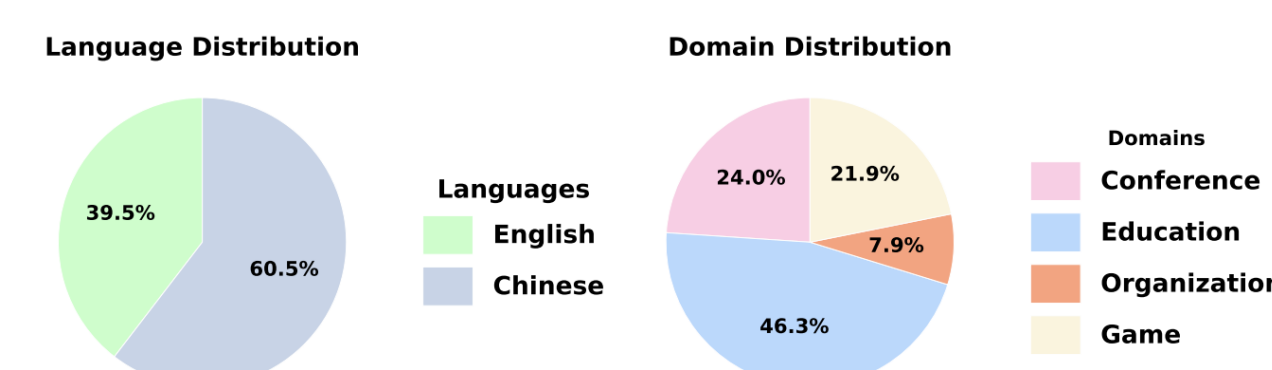
(c) Synthetic QA-Pairs

(d) Verified QA-Pairs

Data Generation Pipeline. We first collect root official websites. Then we mimic human behavior by systematically clicking and collecting subpages accessible through sublinks on the root page. Using predefined rules, we leverage GPT4o to generate synthetic QA-pairs based on the gathered information, followed by manual verification to ensure accuracy and relevance.

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

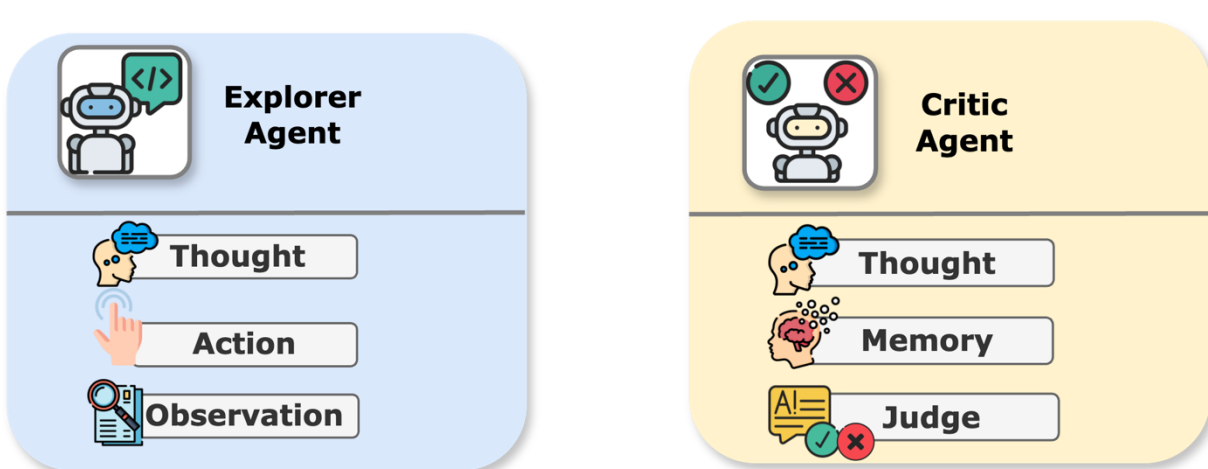
Dataset statistics on **data difficulty level**.



The **language and domain** distribution.

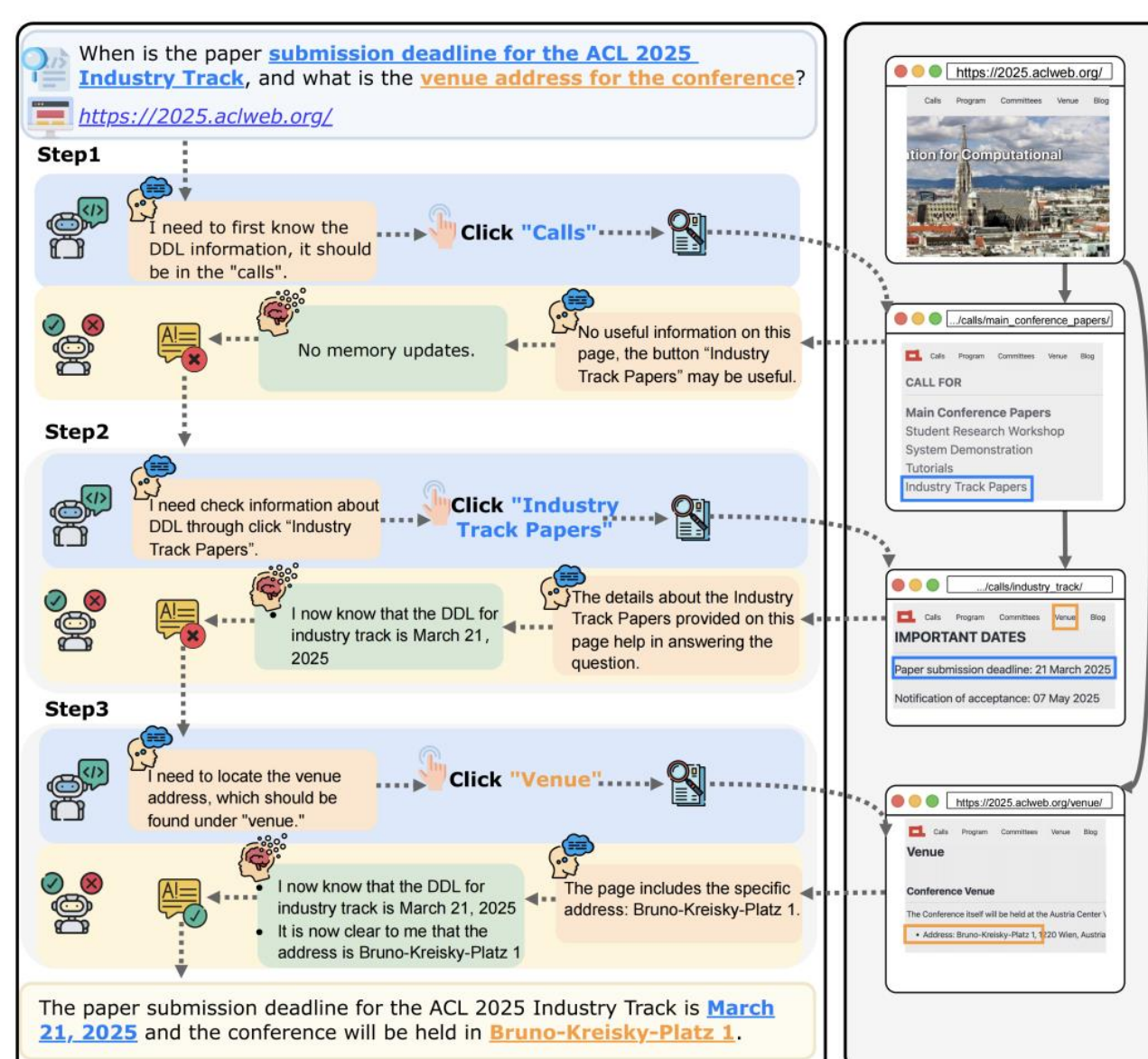
We obtain **680** question-answer pairs for WebWalkerQA.

WebWalker



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** motivated by pair programming.









Overall framework.

		Single-source QA						Multi-source QA						Overall	
Backbones	Method	Easy		Medium		Hard		Easy		Medium		Hard			
		acc.	A.C.	acc.	A.C.	acc.	A.C.	acc.	A.C.	acc.	A.C.	acc.	A.C.		
Closed-Sourced LLMs															
GPT-4o	ReAct	53.75	2.53	45.00	3.34	30.00	5.61	32.50	2.34	31.43	3.97	15.00	6.77	33.82	3.8
	Reflexion WebWalker	56.25	2.91	51.43	3.88	30.83	5.75	35.00	3.67	27.14	4.13	16.67	7.05	35.29	4.2
Qwen-Plus	ReAct	55.00	2.97	50.00	3.43	30.00	6.02	47.50	4.00	34.29	3.85	15.83	6.57	37.50	4.6
	Reflexion WebWalker	48.75	1.67	48.57	2.69	28.33	4.00	35.00	2.60	27.86	3.11	14.17	6.55	33.08	3.0
Open-Sourced LLMs															
Qwen-2.5-7B	ReAct	37.50	3.36	18.57	4.88	9.17	5.45	17.50	3.42	11.43	3.62	5.83	4.57	16.02	2.9
	Reflexion WebWalker	37.50	4.03	25.00	3.48	11.67	4.57	30.00	2.66	15.71	5.45	4.17	7.8	19.11	4.0
Qwen-2.5-14B	ReAct	41.25	3.39	24.71	3.86	12.50	5.93	18.75	3.00	20.71	3.34	5.83	7.28	19.85	3.9
	Reflexion WebWalker	36.25	1.86	32.14	2.75	15.00	3.61	27.50	2.31	22.86	3.00	5.00	5.00	22.35	2.7
Qwen-2.5-32B	ReAct	46.25	2.21	34.29	2.83	15.00	4.44	36.25	2.51	22.86	3.34	5.83	5.42	25.14	3.0
	Reflexion WebWalker	41.25	2.42	41.43	3.24	23.33	4.42	30.00	3.95	22.86	3.56	10.00	6.16	27.50	3.6
Qwen-2.5-72B	ReAct	47.50	2.21	35.71	3.20	16.67	3.55	36.25	2.68	18.57	3.00	8.33	3.70	25.44	2.9
	Reflexion WebWalker	42.50	2.52	32.86	2.65	16.67	3.90	31.25	2.84	23.57	3.12	5.83	5.00	23.26	3.0
Qwen-2.5-72B	ReAct	47.50	1.68	38.57	2.79	20.00	4.04	45.00	2.25	32.14	3.13	10.00	5.41	30.73	2.8
	Reflexion WebWalker	57.50	3.04	44.29	3.88	28.33	5.82	36.25	3.62	25.00	3.60	12.50	6.26	32.50	4.0
		58.75	2.70	48.57	3.07	25.83	5.77	35.00	3.57	29.19	4.87	15.00	5.38	33.26	4.3

Benchmark results across closed-sourced and open-sourced LLMs as the backbone. *Acc.* and *A.C.* refer to accuracy and action count, respectively.

- Larger LLMs have enhanced **long-range information-seeking ability**.
- Even the best WebWalker with GPT-4o scores under 40%, underscoring difficulty.
- As depth or source count increases, acquiring the needed information becomes harder.

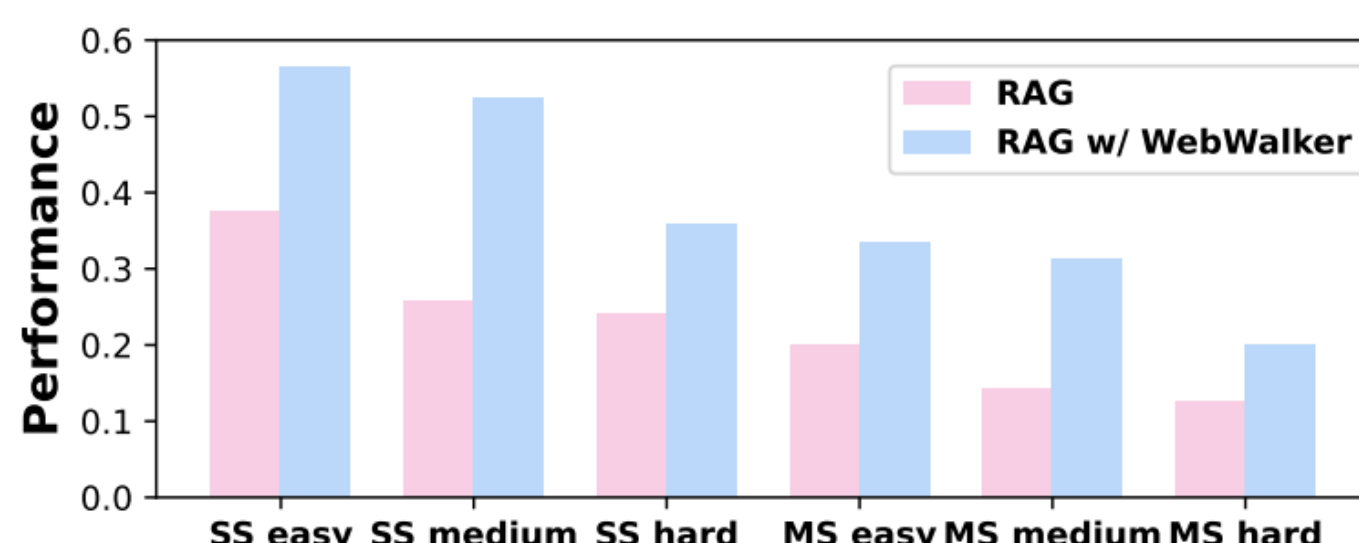
Findings 1

Systems	Single-source QA			Multi-source QA			Overall
							
	Easy	Medium	Hard	Easy	Medium	Hard	
Close Book (No Retrieval)							
Gemini-1.5-Pro o1-preview	12.50	7.86	8.33	11.25	6.43	5.00	8.08
	16.25	10.00	9.17	7.50	10.71	6.67	9.85
Commercial Systems							
Doubao	45.00	15.00	18.33	13.75	8.57	10.00	16.76
Gemini-Search	40.00	32.14	29.17	30.00	23.57	17.50	27.94
ERNIE-4.0-8K	52.50	30.00	28.33	21.25	18.57	30.00	28.97
Kimi	77.50	41.43	40.83	26.25	26.43	22.50	37.35
Tongyi	41.25	45.00	41.67	40.00	41.43	34.17	40.73
Open-Sourced Systems							
Naive RAG	37.50	25.71	24.17	20.00	14.29	12.50	20.73
MindSearch	15.00	11.43	10.83	8.75	12.14	10.00	11.32
Avg.	37.50	24.29	23.42	19.86	18.02	16.48	-

Results on Commercial and Open-sourced Searched-enhanced RAG systems.

Findings (i): RAG systems struggle with key challenges that require effective web traversal.

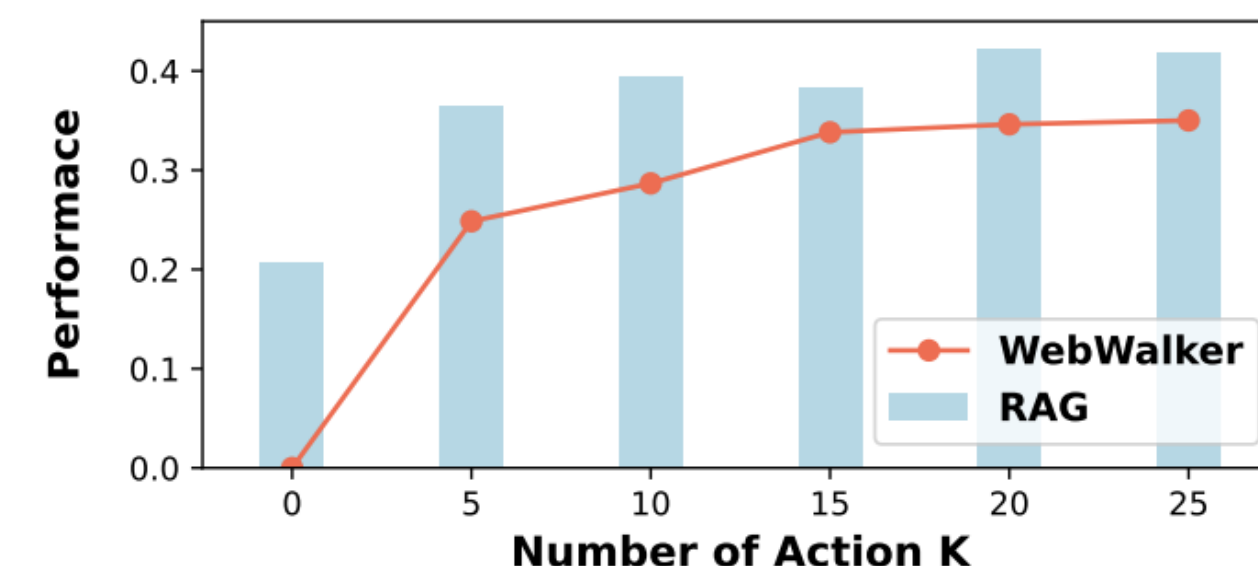
Findings 2



Performance under standard RAG and **RAG combined with WebWalker configurations**,

Findings (ii): WebWalker can be a module in agentic RAG system, enabling vertical exploration.

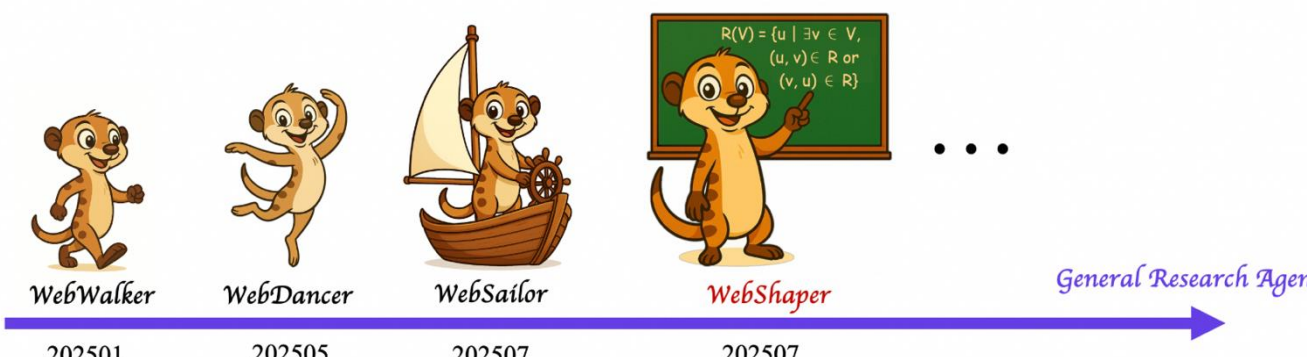
Findings 3



Overall performance on WebWalker and RAG combined WebWalker at **varying values of K**.

Findings (iii): Scaling the process of digging through links could represent a potential direction for vertical exploration in RAG systems.

WebAgent for Information Seeking



Web Agents are **autonomous** systems that perceive their real-world web environment, make decisions, and take actions to accomplish specific and human-like tasks.

If you like our project, feel free to give us a ★ on GitHub©