TONGYI 人人成长 Happy Growth

Mar

PAPER READING

Please search ATA-Paper Reading program for more information

通义

WebWalker: Benchmarking LLMs in Web Traversal

Jialong Wu, Wenbiao Yin, Jiang Yong, Zhenglin Wang, Zekun Xi, Runnan Fang
Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, Fei Huang

ACL 2025 Submission

吴家隆 （承桦 ）| Research Intern | 通义实验室-自然语言智能 | 主管: 咏江

content

目录_

01 / **Background and Motivation**

Brief History of Web Agents and RAG Limitation

02 / **Dataset and Method**

Introduce WebWalkerQA and WebWalker

03 / **Result and Discussion**

Results on Agents and RAG systems

04 / **Conclusion**

Limitations and Future Works

01 /

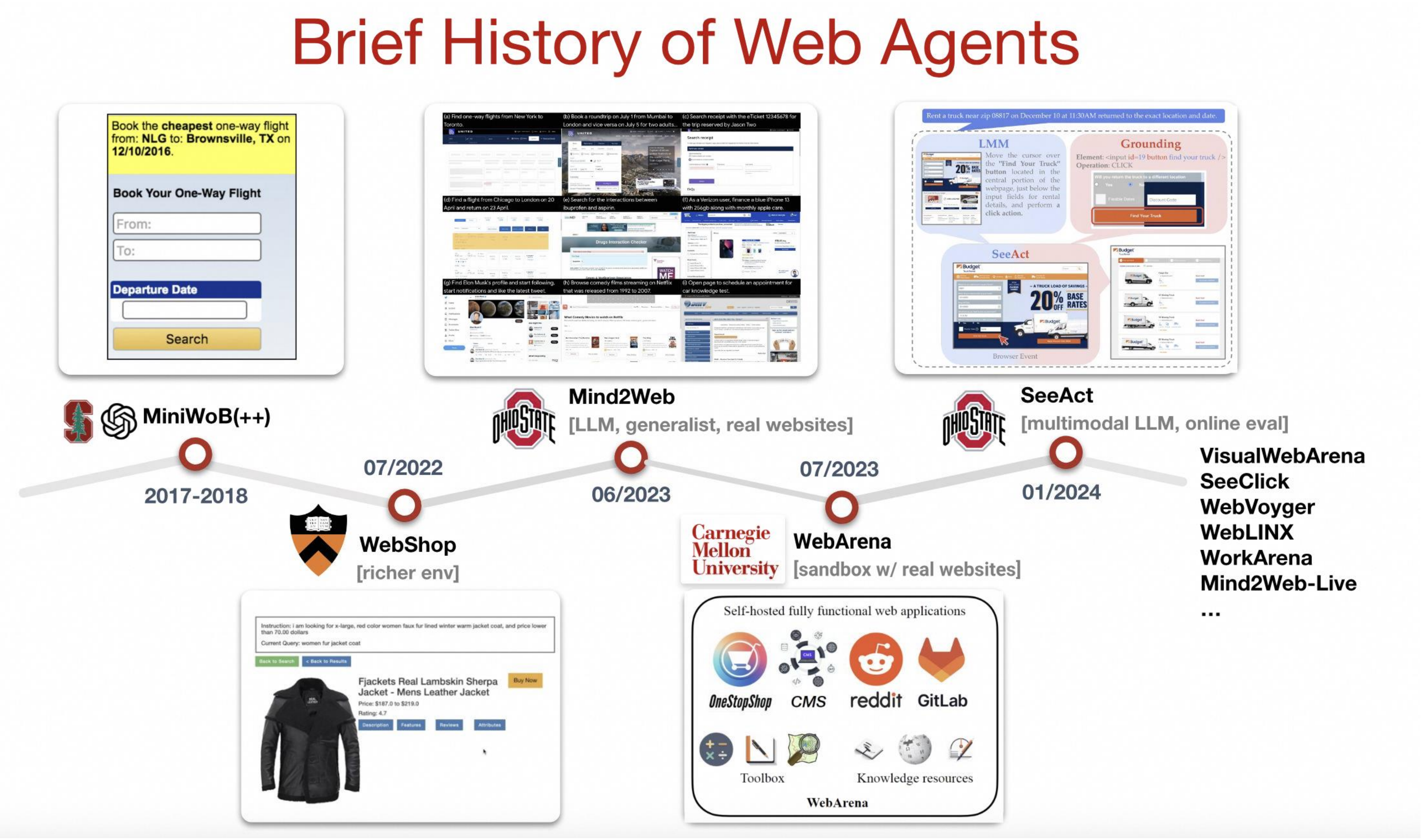
BACKGROUND AND MOTIVATION

▼

Brief History of Web Agents and RAG Limitation

BACKGROUND AND MOTIVATION

Brief History of Web Agents and RAG Limitation



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]

Instruction: i am looking for x-large, red color women faux fur lined winter warm jacket coat, and price lower than 70.00 dollars

Current Query: women fur jacket coat

Back to Search | < Back to Results

Fjackets Real Lambskin Sherpa Jacket - Mens Leather Jacket

Buy Now

Price: $187.0 to $219.0

Rating: 4.7

Description | Features | Reviews | Attributes

(a) Find one-way flights from New York to Toronto.

(b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults...

(c) Search receipt with the eTicket 12345678 for the trip reserved by Jason Two

(d) Find a flight from Chicago to London on 20 April and return on 23 April.

(e) Search for the interactions between ibuprofen and aspirin.

(f) As a Verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.

(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet.

(h) Browse comedy films streaming on Netflix that was released from 1992 to 2007.

(i) Open page to schedule an appointment for car knowledge test.

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

Rent a truck near zip 08817 on December 10 at 11:30AM returned to the exact location and date.

LMM

Move the cursor over the **"Find Your Truck" button** located in the central portion of the webpage, just below the input fields for rental details, and perform **a click action.**

Grounding

Element: <input id=19 button find your truck />
Operation: CLICK

Will you return the truck to a different location

Find Your Truck

SeeAct

Browser Event

SeeAct

[multimodal LLM, online eval]

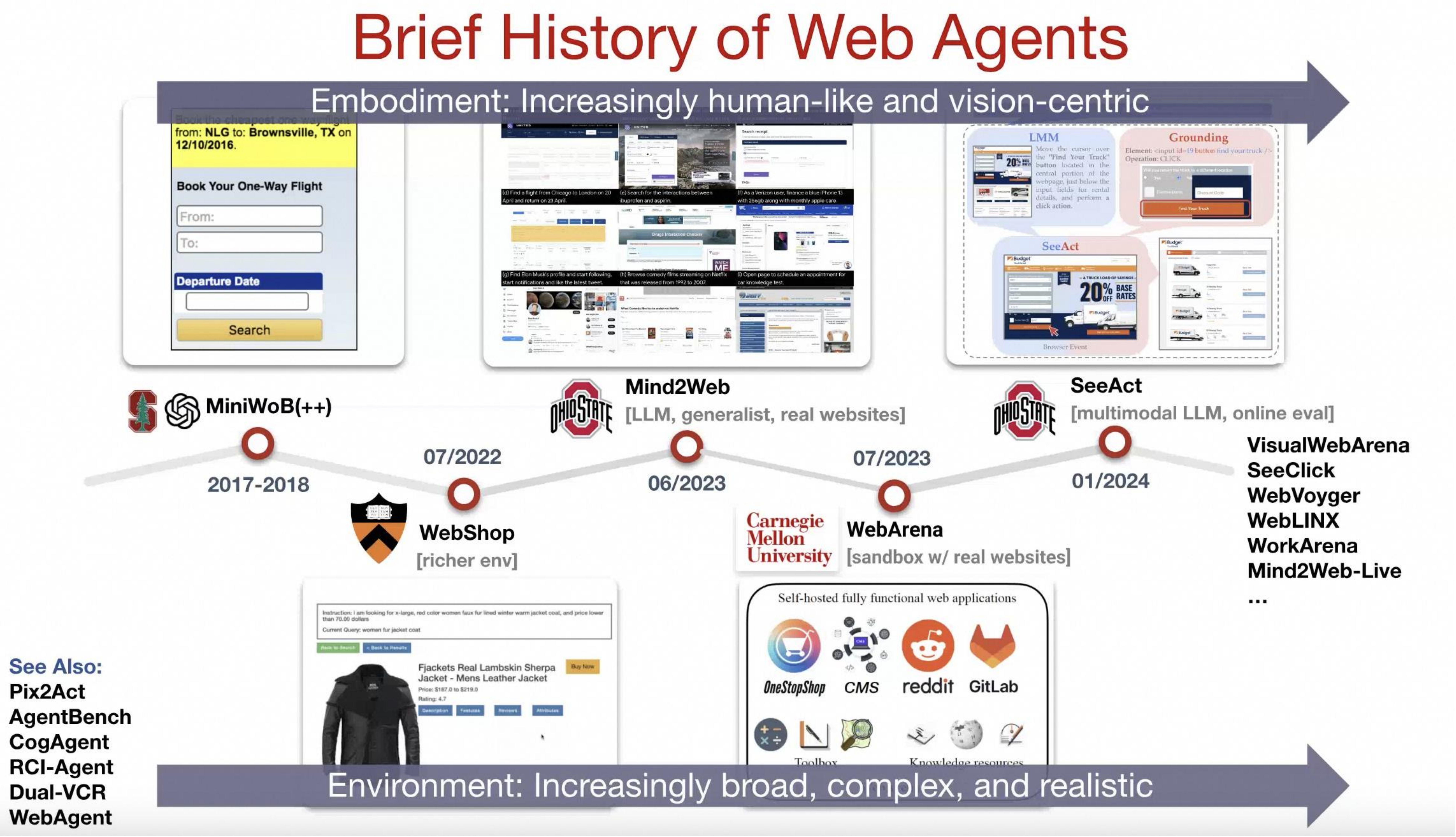
01/2024

VisualWebArena
SeeClick
WebVoyger
WebLINX
WorkArena
Mind2Web-Live
...

From SPNLP@ACL2024 Workshop Tutorial

BACKGROUND AND MOTIVATION

Brief History of Web Agents and RAG Limitation



Brief History of Web Agents

Embodiment: Increasingly human-like and vision-centric

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

(d) Find a flight from Chicago to London on 20 April and return on 23 April.

(e) Search for the interactions between ibuprofen and aspirin.

(f) As a Verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.

(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet.

(h) Browse comedy films streaming on Netflix that was released from 1992 to 2007.

(i) Open page to schedule an appointment for car knowledge test.

LMM: Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.

Grounding: Element: <input id=19 button find your truck /> Operation: CLICK

MiniWoB(++)
2017-2018

07/2022
WebShop
[richer env]

Mind2Web
[LLM, generalist, real websites]
06/2023

07/2023
WebArena
[sandbox w/ real websites]

SeeAct
[multimodal LLM, online eval]
01/2024

VisualWebArena
SeeClick
WebVoyger
WebLINX
WorkArena
Mind2Web-Live
...

Instruction: I am looking for x-large, red color women faux fur lined winter warm jacket coat, and price lower than 70.00 dollars

Current Query: women fur jacket coat

Self-hosted fully functional web applications

See Also:
Pix2Act
AgentBench
CogAgent
RCI-Agent
Dual-VCR
WebAgent

Environment: Increasingly broad, complex, and realistic

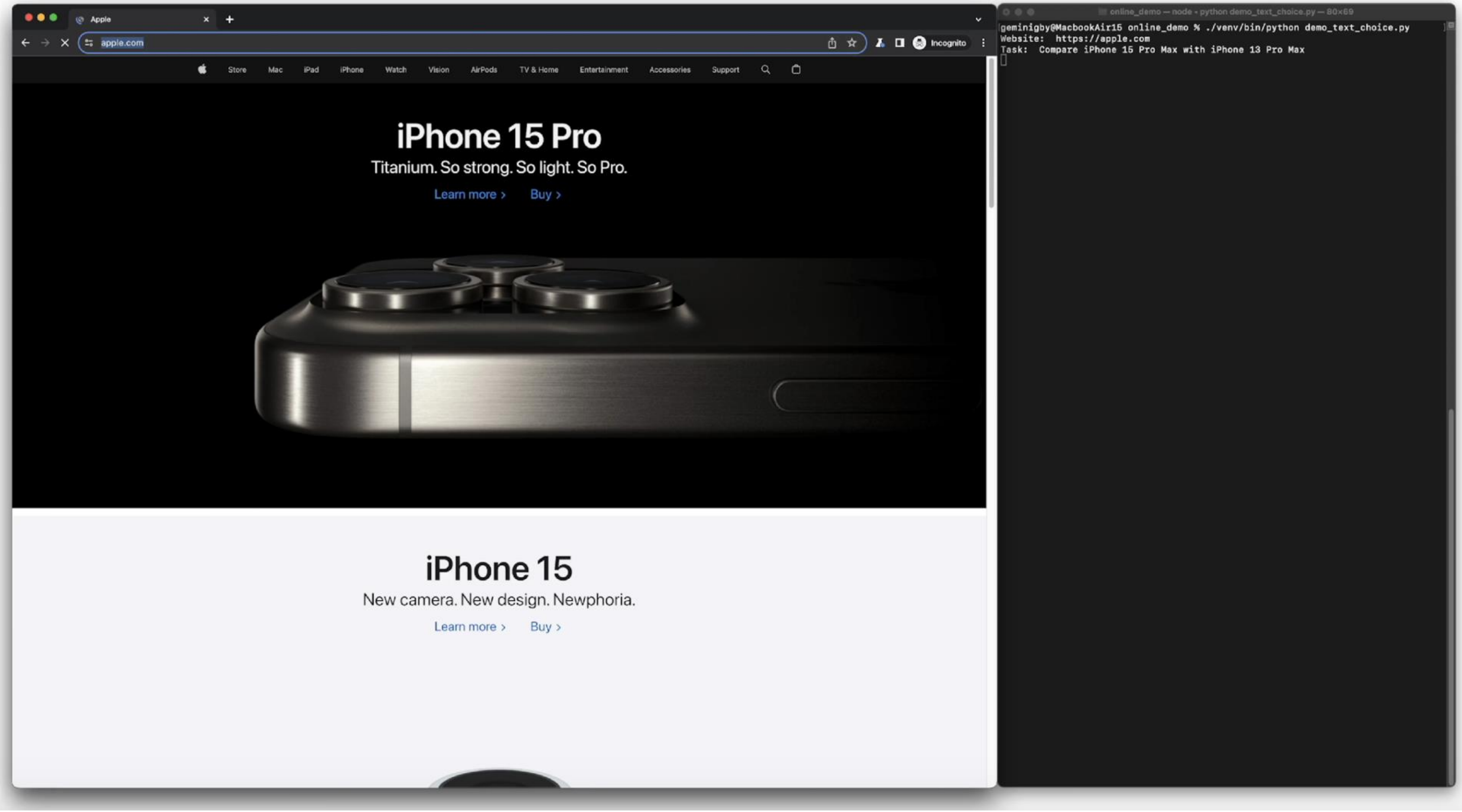
From SPNLP@ACL2024 Workshop Tutorial

BACKGROUND AND MOTIVATION

Brief History of Web Agents and RAG Limitation

(NeurIPS'23) (ICML'24)

Generalist Web Agents: Mind2Web & SeeAct



```
geminigby@MacbookAir15 online_demo % ./venv/bin/python demo_text_choice.py
Website:  https://apple.com
Task:  Compare iPhone 15 Pro Max with iPhone 13 Pro Max
```

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

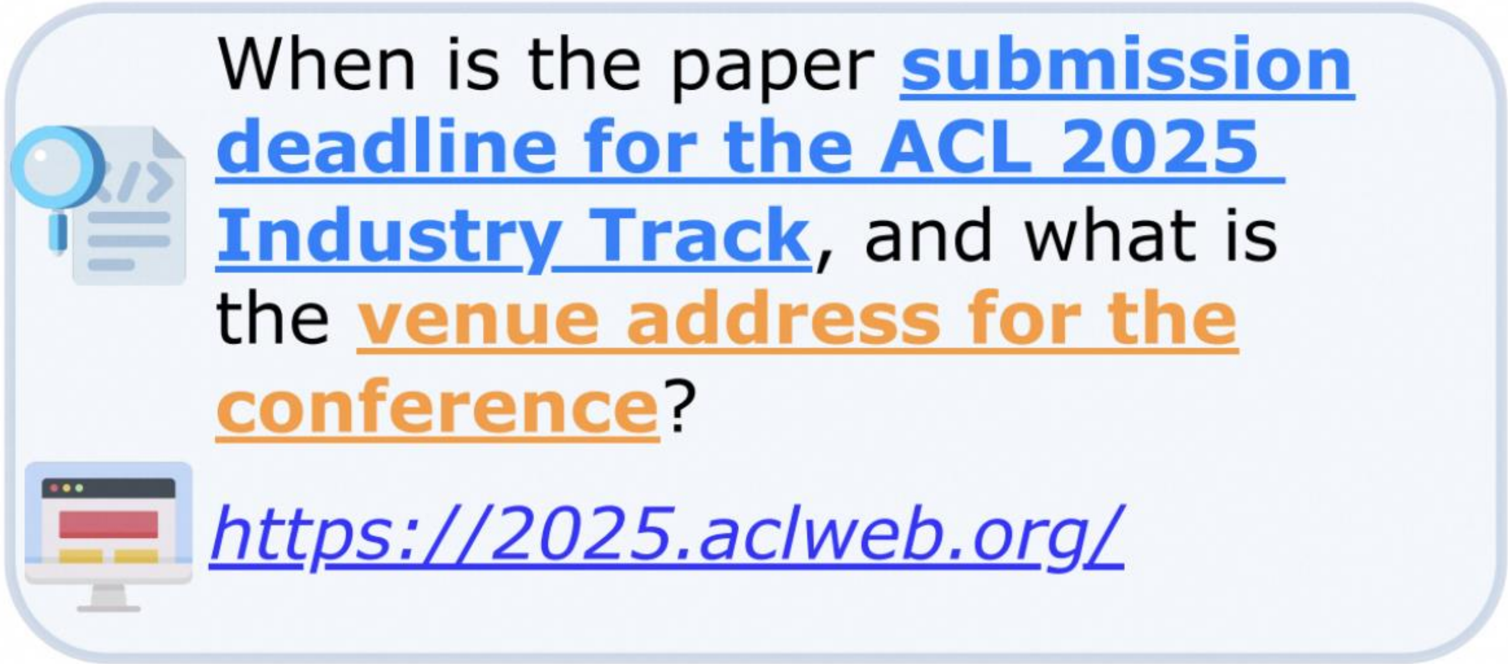
From SPNLP@ACL2024
Workshop Tutorial

BACKGROUND AND MOTIVATION

Brief History of Web Agents and RAG Limitation

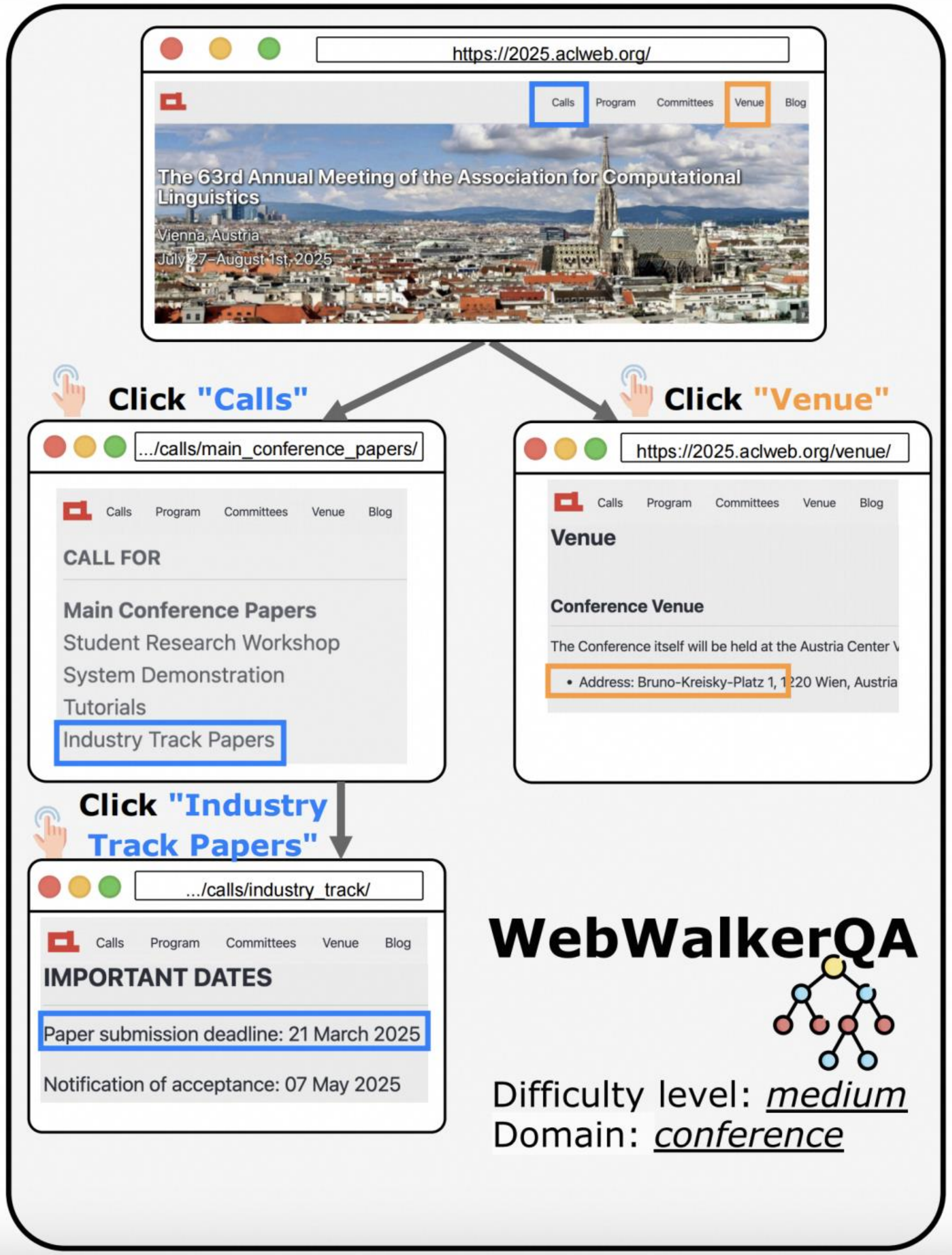
Key challenge in RAG:

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/



https://2025.aclweb.org/

Calls Program Committees Venue Blog

The 63rd Annual Meeting of the Association for Computational Linguistics

Vienna, Austria

July 27–August 1st, 2025

Click "Calls"

.../calls/main_conference_papers/

Calls Program Committees Venue Blog

CALL FOR

Main Conference Papers

Student Research Workshop

System Demonstration

Tutorials

Industry Track Papers

Click "Venue"

https://2025.aclweb.org/venue/

Calls Program Committees Venue Blog

Venue

Conference Venue

The Conference itself will be held at the Austria Center V

- Address: Bruno-Kreisky-Platz 1, 1220 Wien, Austria

Click "Industry Track Papers"

.../calls/industry_track/

Calls Program Committees Venue Blog

IMPORTANT DATES

Paper submission deadline: 21 March 2025

Notification of acceptance: 07 May 2025

WebWalkerQA

Difficulty level: *medium*

Domain: *conference*

BACKGROUND AND MOTIVATION

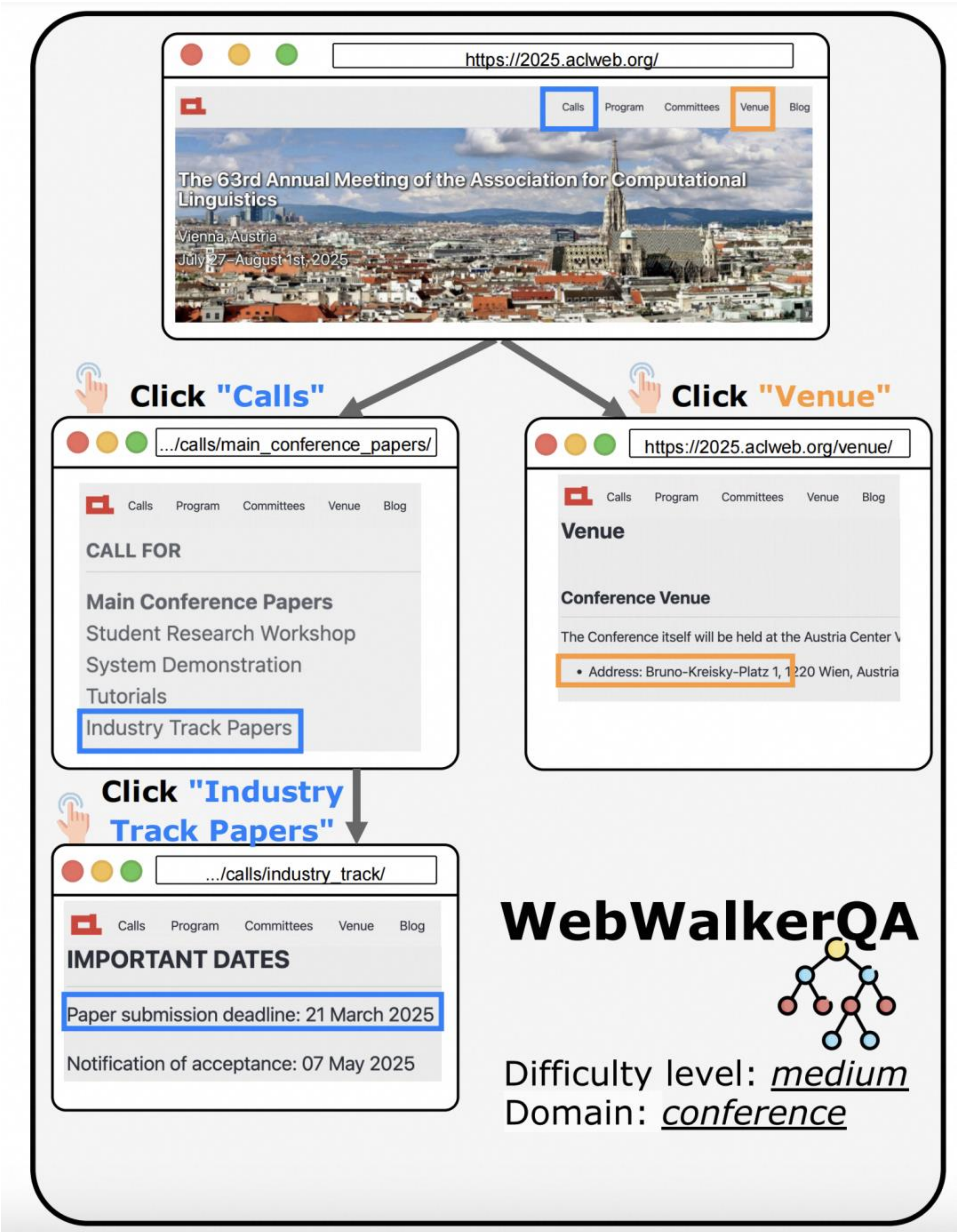
Brief History of Web Agents and RAG Limitation

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.

- 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**.



https://2025.aclweb.org/

The 63rd Annual Meeting of the Association for Computational Linguistics

Vienna, Austria

July 27–August 1st, 2025

Click "Calls"

.../calls/main_conference_papers/

CALL FOR

Main Conference Papers

Student Research Workshop

System Demonstration

Tutorials

Industry Track Papers

Click "Venue"

https://2025.aclweb.org/venue/

Venue

Conference Venue

The Conference itself will be held at the Austria Center V

- Address: Bruno-Kreisky-Platz 1, 1220 Wien, Austria

Click "Industry Track Papers"

.../calls/industry_track/

IMPORTANT DATES

Paper submission deadline: 21 March 2025

Notification of acceptance: 07 May 2025

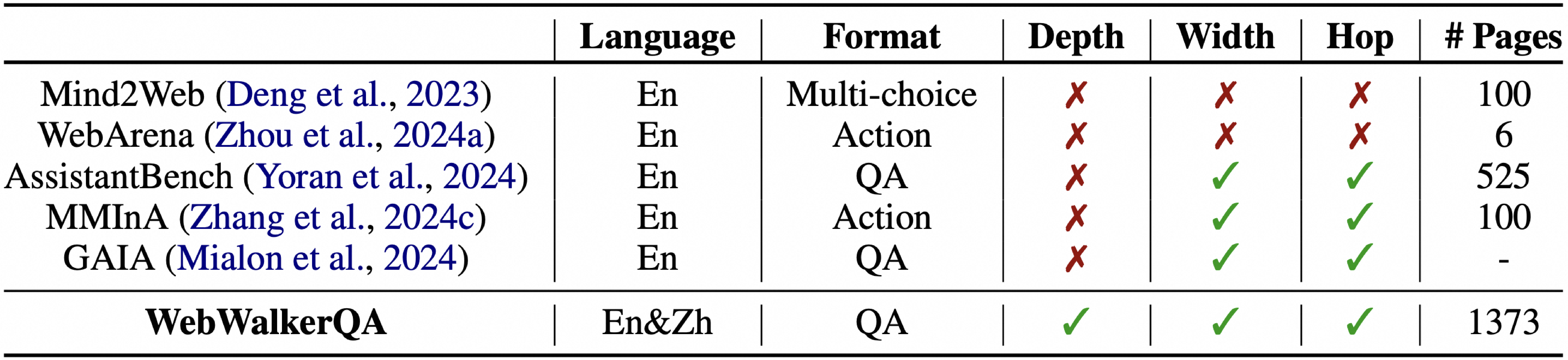
WebWalkerQA

Difficulty level: *medium*

Domain: *conference*

DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker



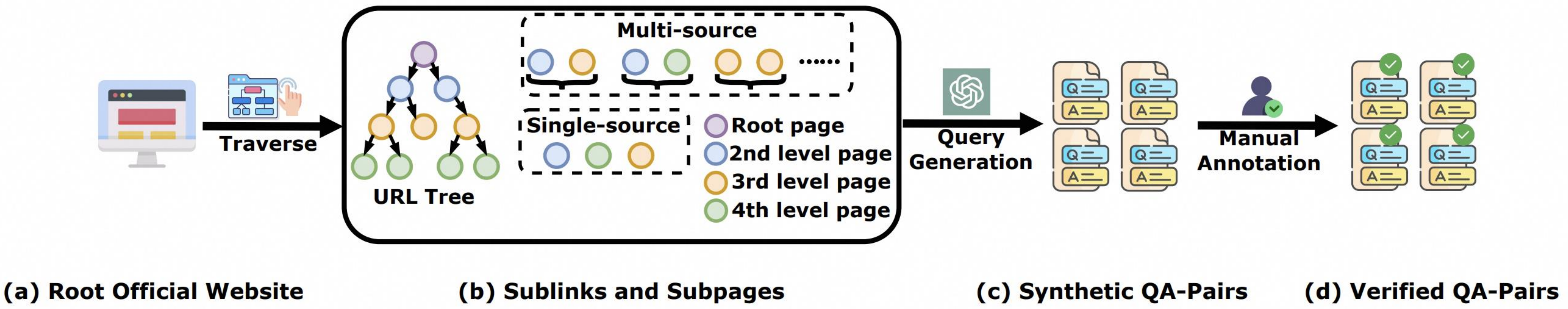
	Language	Format	Depth	Width	Hop	# Pages
Mind2Web (Deng et al., 2023)	En	Multi-choice	✗	✗	✗	100
WebArena (Zhou et al., 2024a)	En	Action	✗	✗	✗	6
AssistantBench (Yoran et al., 2024)	En	QA	✗	✓	✓	525
MMInA (Zhang et al., 2024c)	En	Action	✗	✓	✓	100
GAIA (Mialon et al., 2024)	En	QA	✗	✓	✓	-
WebWalkerQA	En&Zh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

Comparison between WebWalkerQA and other benchmarks.

DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker

WebWalkerQA

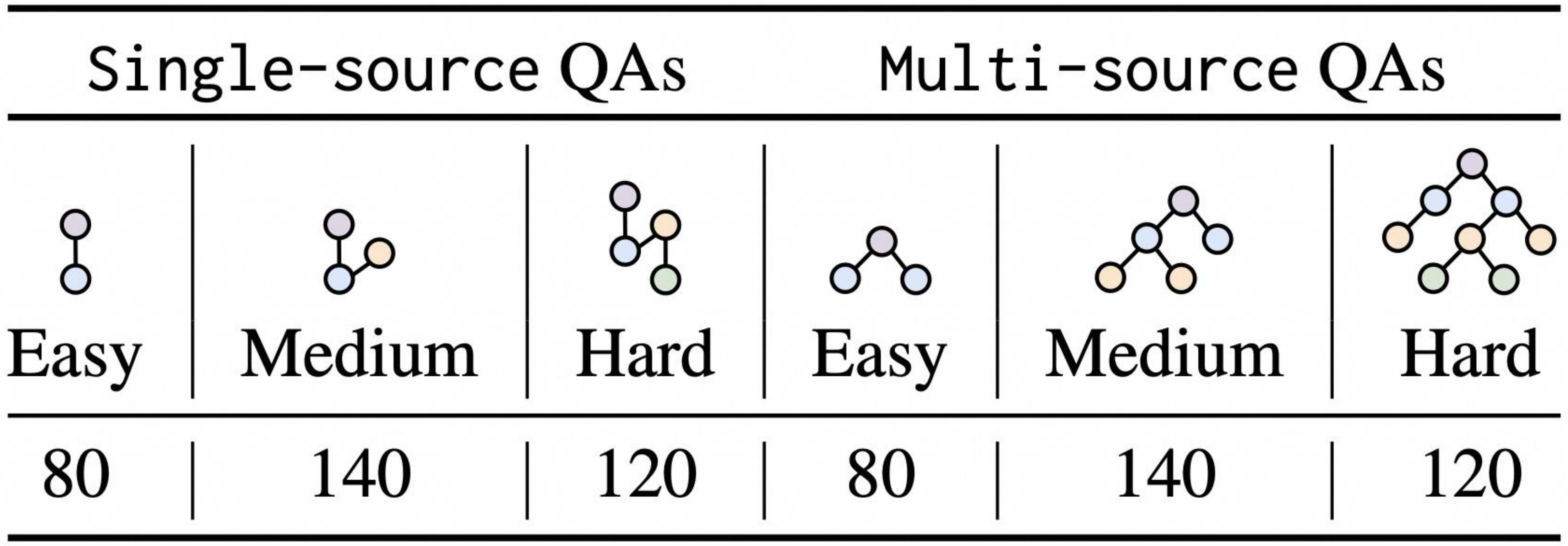


Data Generation Pipeline for WebWalkerQA.

WebWalkerQA

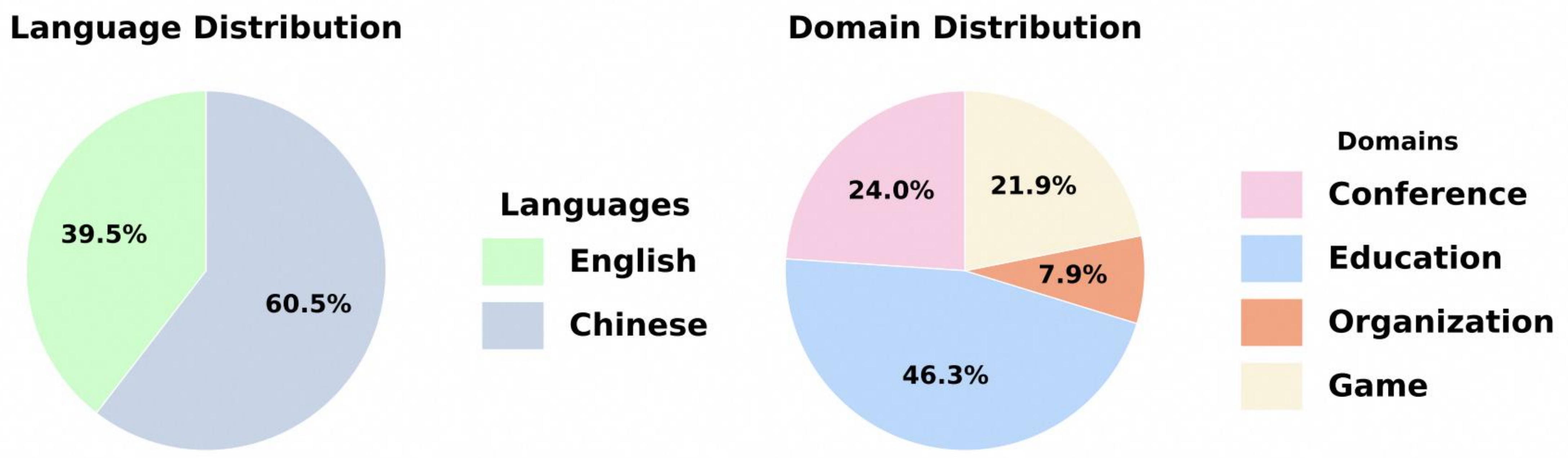
DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker



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 AND METHODS

Introduce WebWalkerQA and WebWalker

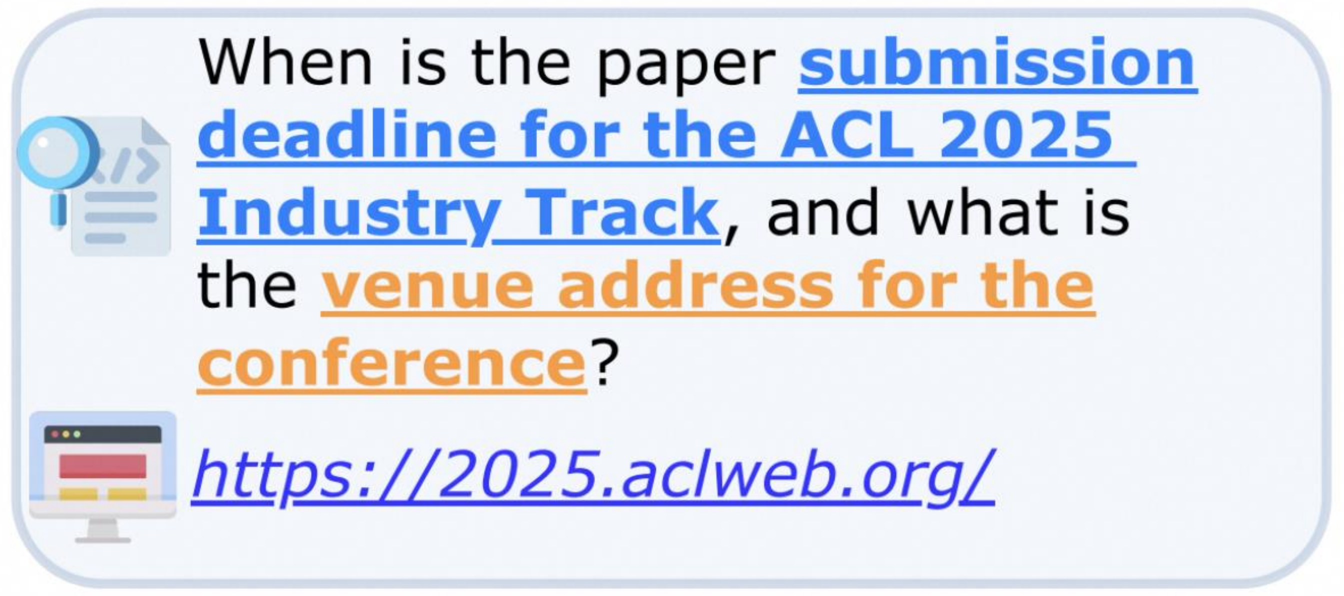
Web Traversal Task:

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:

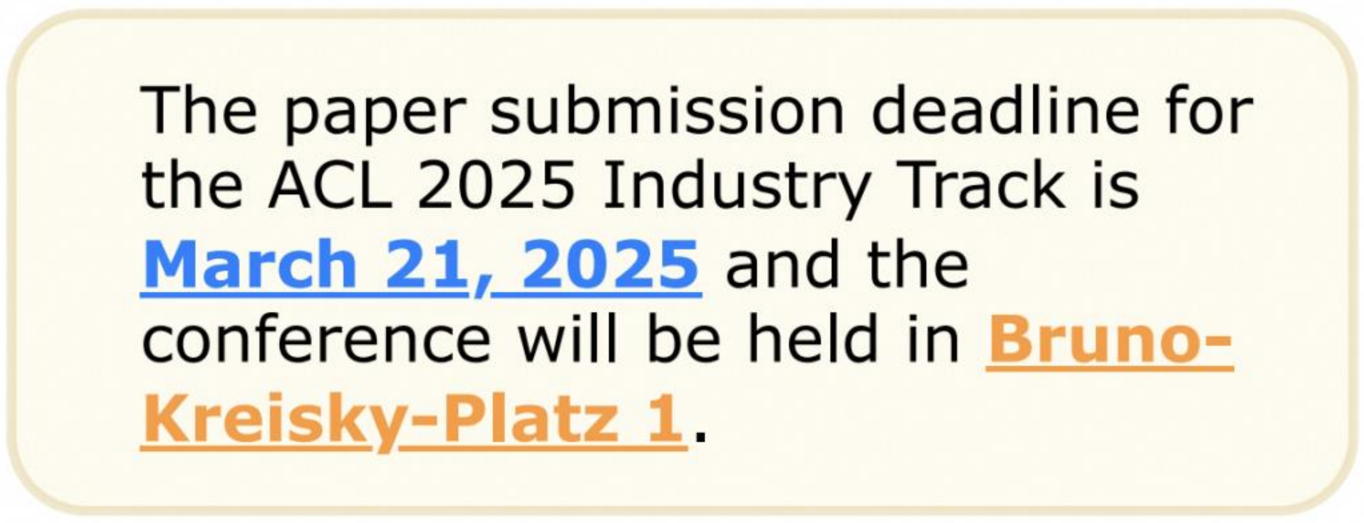
Correctness -> *acc.* Evaluated by GPT-4o

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**.

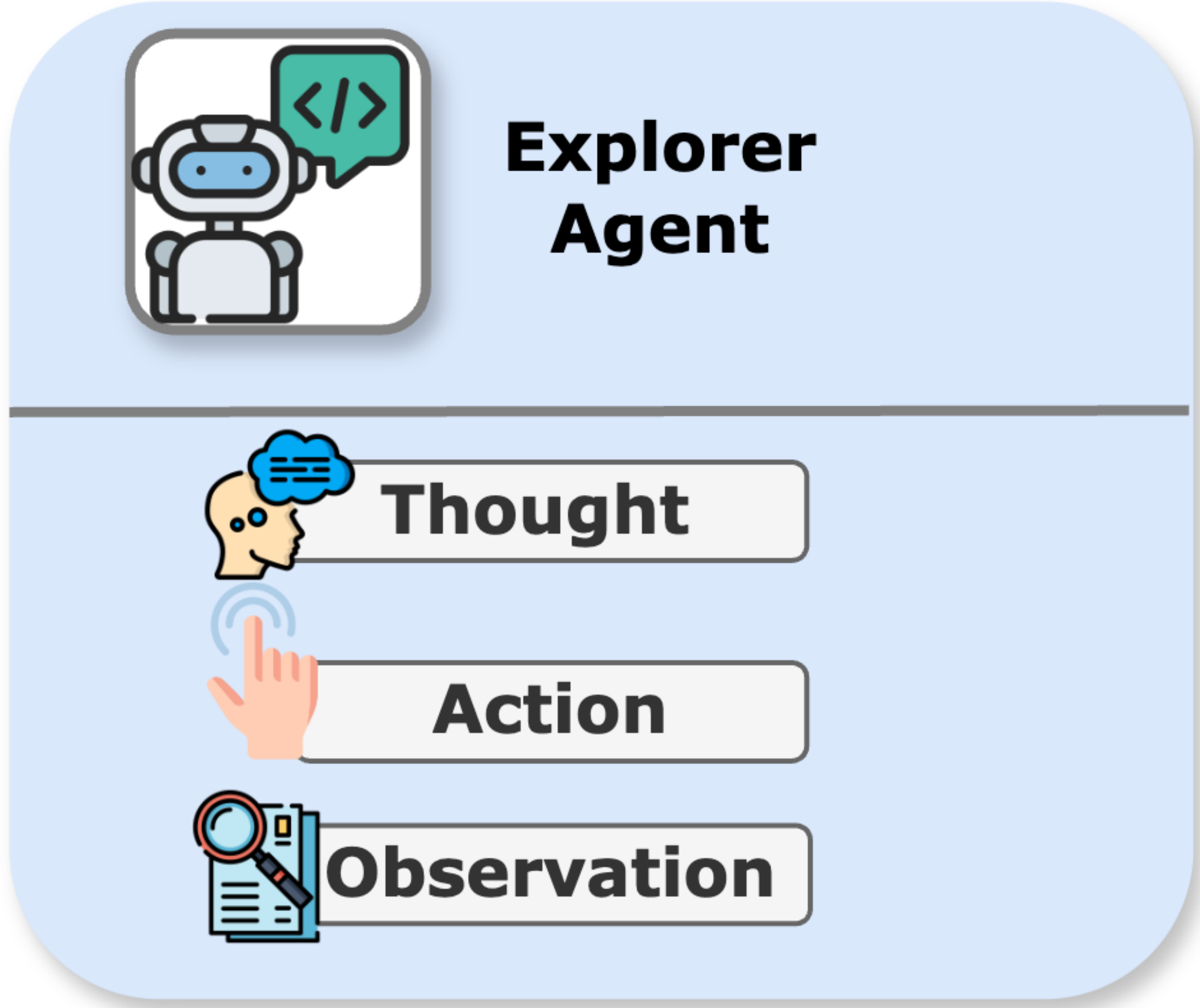
DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker

Explorer Agent + Critic Agent

WebWalker

WebWalker: a multi-agent framework



Explorer Agent

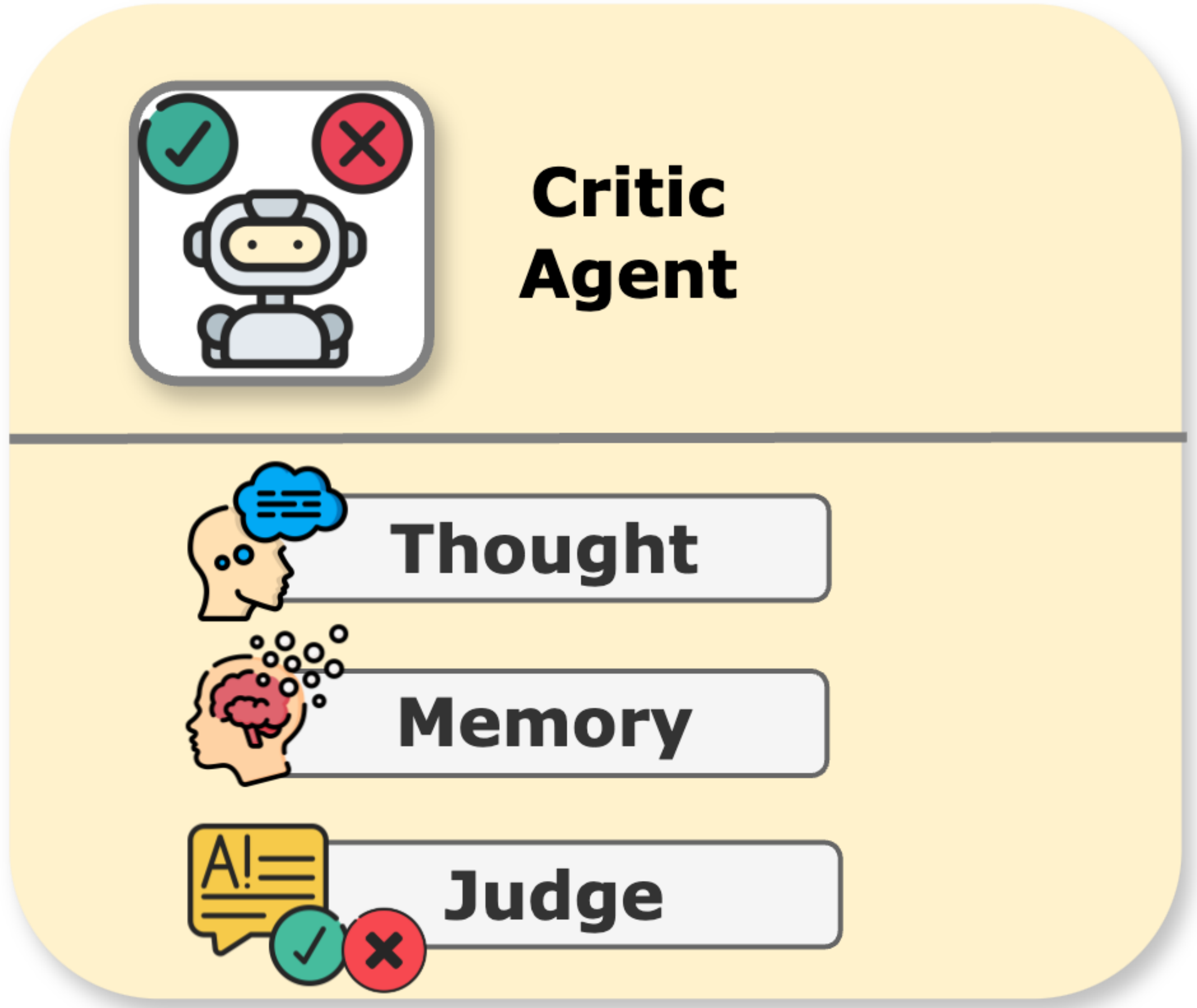
Thought

Action

Observation

Think then Explore

ReAct format



Critic Agent

Thought

Memory

Judge

Think then Critique

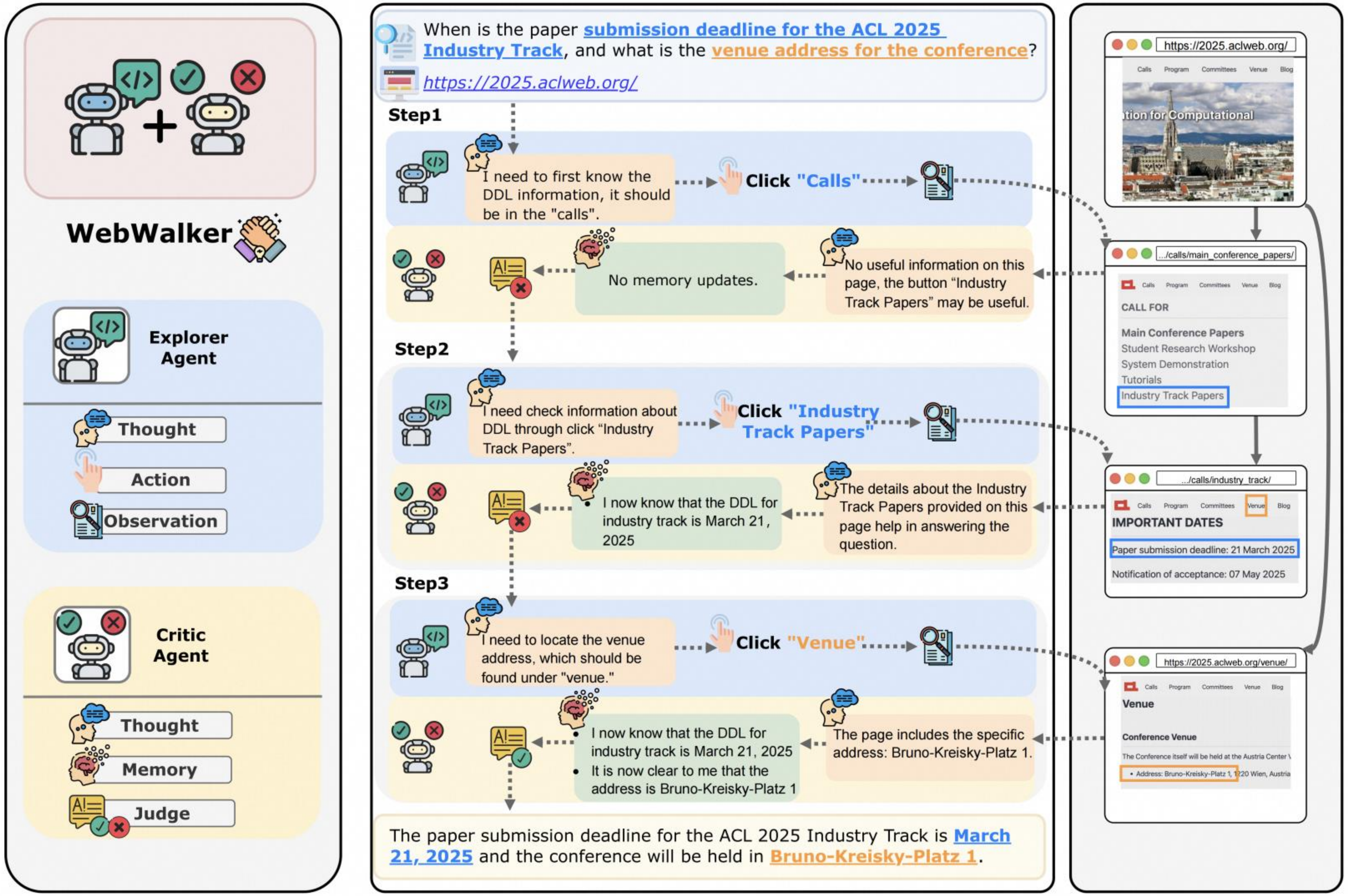
Motivated by pair programming

DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker

Explorer Agent + Critic Agent

WebWalker



WebWalker

Explorer Agent

- Thought
- Action
- Observation

Critic Agent

- Thought
- Memory
- Judge

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/

Step1

I need to first know the DDL information, it should be in the "calls".

Click "Calls"

No useful information on this page, the button "Industry Track Papers" may be useful.

No memory updates.

Step2

I need check information about DDL through click "Industry Track Papers".

Click "Industry Track Papers"

The details about the Industry Track Papers provided on this page help in answering the question.

- I now know that the DDL for industry track is March 21, 2025

Step3

I need to locate the venue address, which should be found under "venue."

Click "Venue"

The page includes the specific address: Bruno-Kreisky-Platz 1.

- I now know that the DDL for industry track is March 21, 2025
- It is now clear to me that the address is Bruno-Kreisky-Platz 1

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**.

https://2025.aclweb.org/

.../calls/main_conference_papers/

CALL FOR

Main Conference Papers
Student Research Workshop
System Demonstration
Tutorials
Industry Track Papers

.../calls/industry_track/

IMPORTANT DATES

Paper submission deadline: 21 March 2025

Notification of acceptance: 07 May 2025

https://2025.aclweb.org/venue/

Venue

Conference Venue

The Conference itself will be held at the Austria Center \

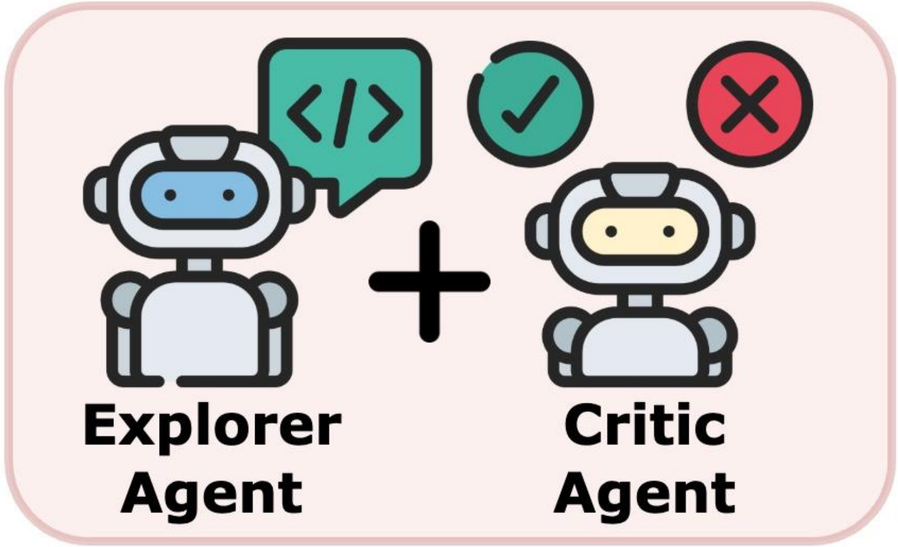
- Address: Bruno-Kreisky-Platz 1, 1220 Wien, Austria

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

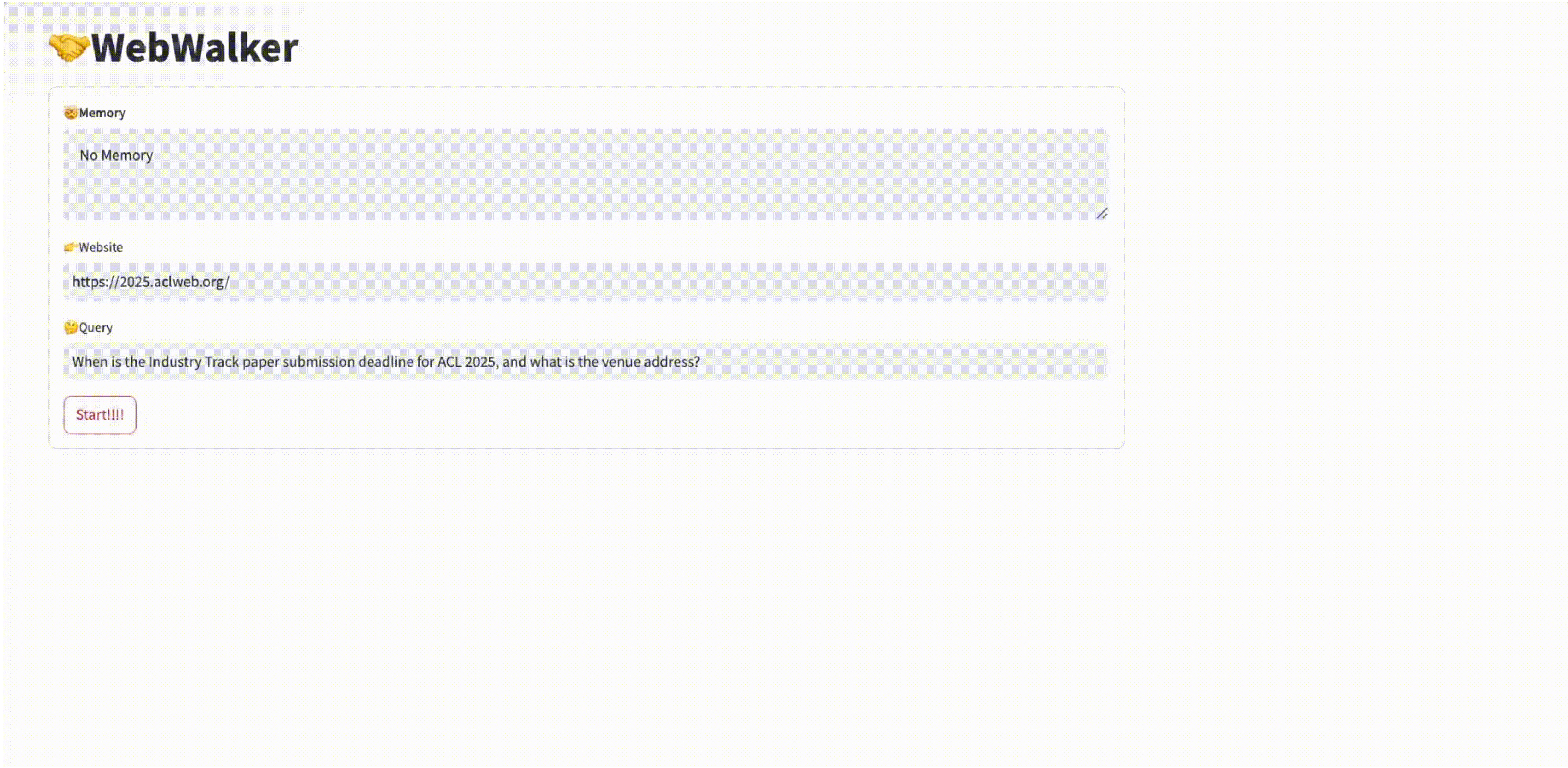
The critic agent **updates the memory** until sufficient information is accumulated to effectively **address the query**.

DATASETS AND METHODS

Introduce WebWalkerQA and WebWalker



WebWalker



🤝WebWalker

Memory

No Memory

Website

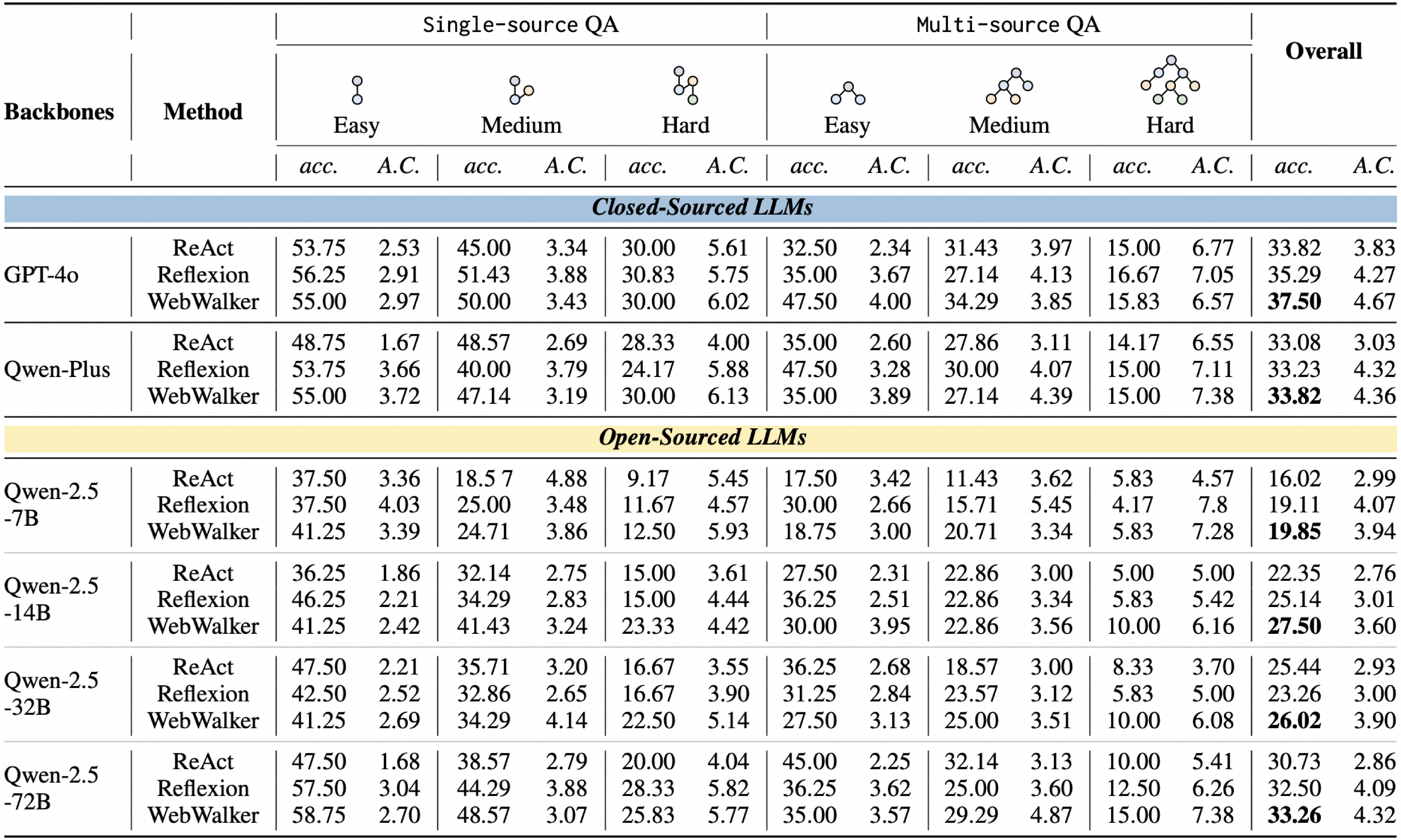
https://2025.aclweb.org/

Query

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

RESULTS AND DISCUSSION

Results on Agents and RAG systems



Backbones	Method	Single-source QA						Multi-source QA						Overall	
		Easy		Medium		Hard		Easy		Medium		Hard			
		acc.	*A.C.*	*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	3.83
	Reflexion	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	4.27
	WebWalker	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**	4.67
Qwen-Plus	ReAct	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	3.03
	Reflexion	53.75	3.66	40.00	3.79	24.17	5.88	47.50	3.28	30.00	4.07	15.00	7.11	33.23	4.32
	WebWalker	55.00	3.72	47.14	3.19	30.00	6.13	35.00	3.89	27.14	4.39	15.00	7.38	**33.82**	4.36
Open-Sourced LLMs															
Qwen-2.5 -7B	ReAct	37.50	3.36	18.5 7	4.88	9.17	5.45	17.50	3.42	11.43	3.62	5.83	4.57	16.02	2.99
	Reflexion	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	4.07
	WebWalker	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**	3.94
Qwen-2.5 -14B	ReAct	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	2.76
	Reflexion	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	3.01
	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**	3.60
Qwen-2.5 -32B	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	2.93
	Reflexion	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	3.00
	WebWalker	41.25	2.69	34.29	4.14	22.50	5.14	27.50	3.13	25.00	3.51	10.00	6.08	**26.02**	3.90
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	2.86
	Reflexion	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	4.09
	WebWalker	58.75	2.70	48.57	3.07	25.83	5.77	35.00	3.57	29.29	4.87	15.00	7.38	**33.26**	4.32

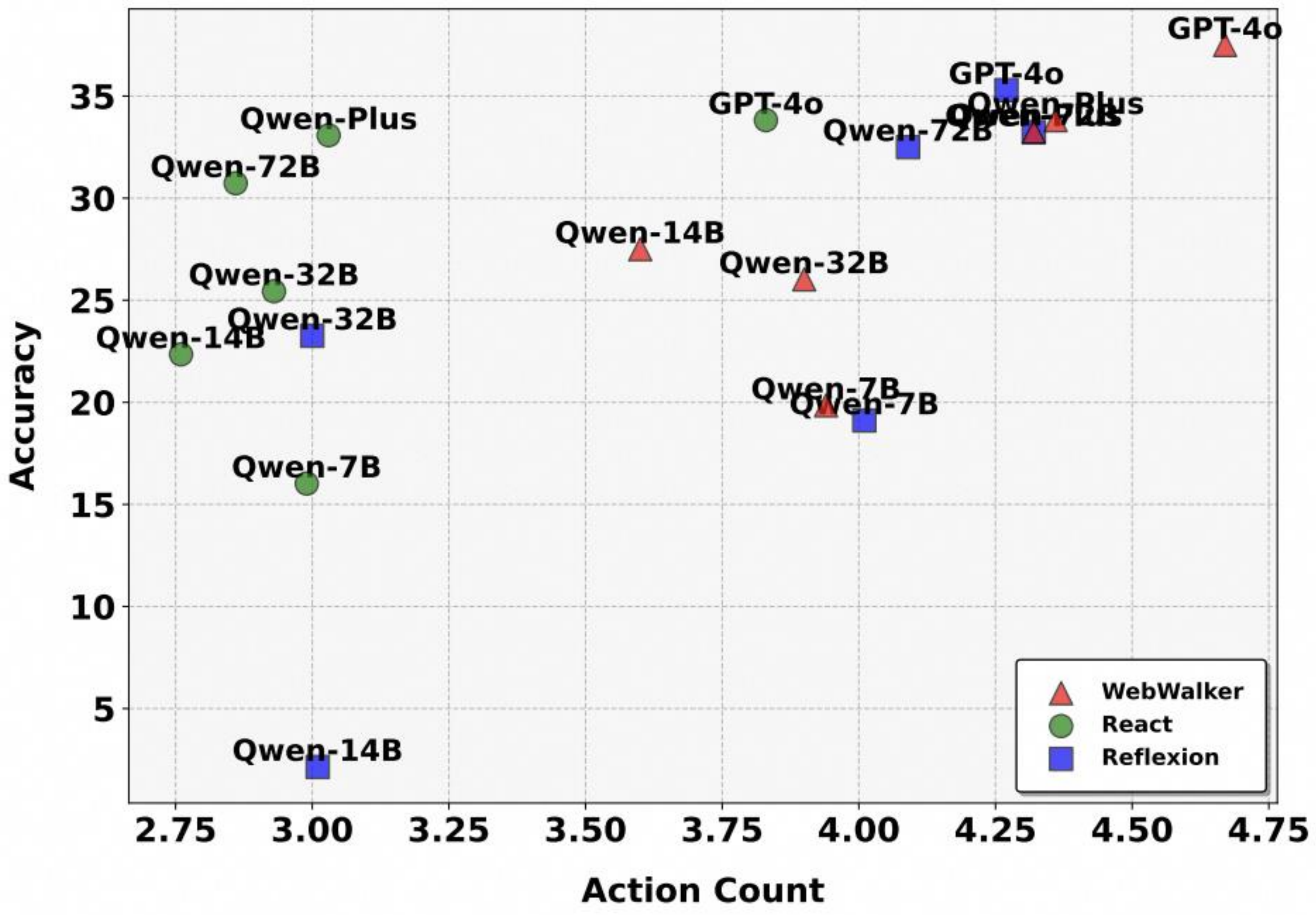
Main results of Agents on WebWalkerQA.

RESULTS AND DISCUSSION

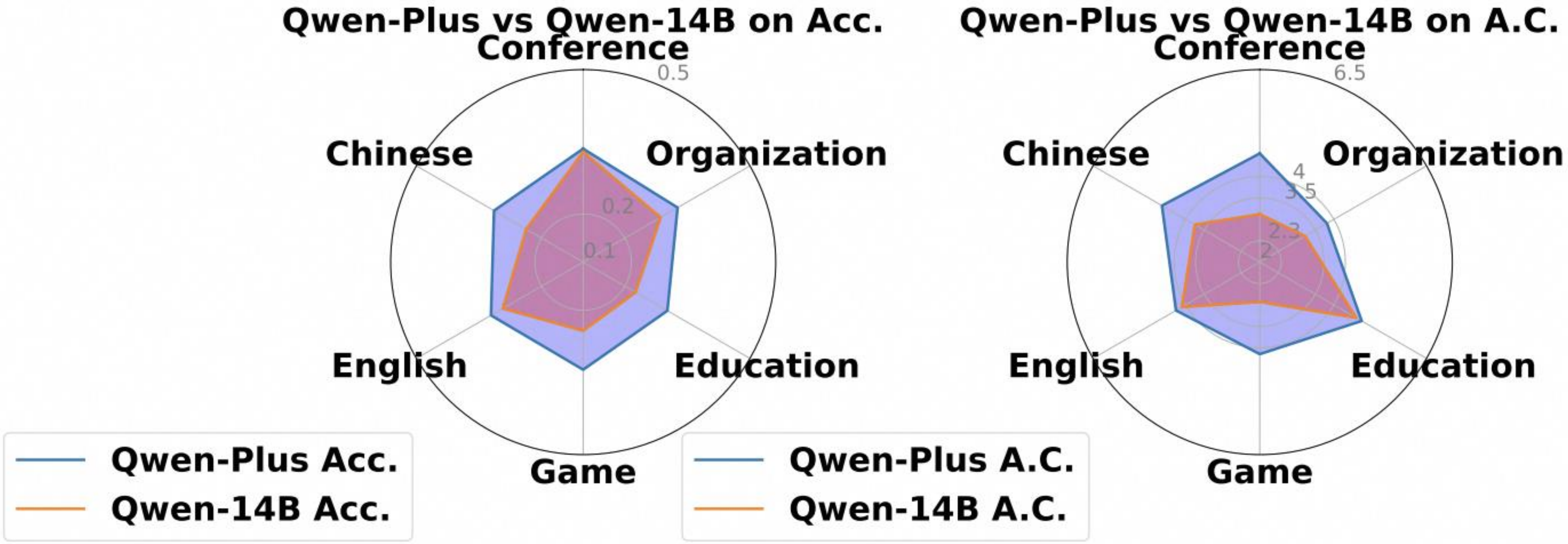
Results on Agents and RAG systems

Explorer Agent + Critic Agent

WebWalker



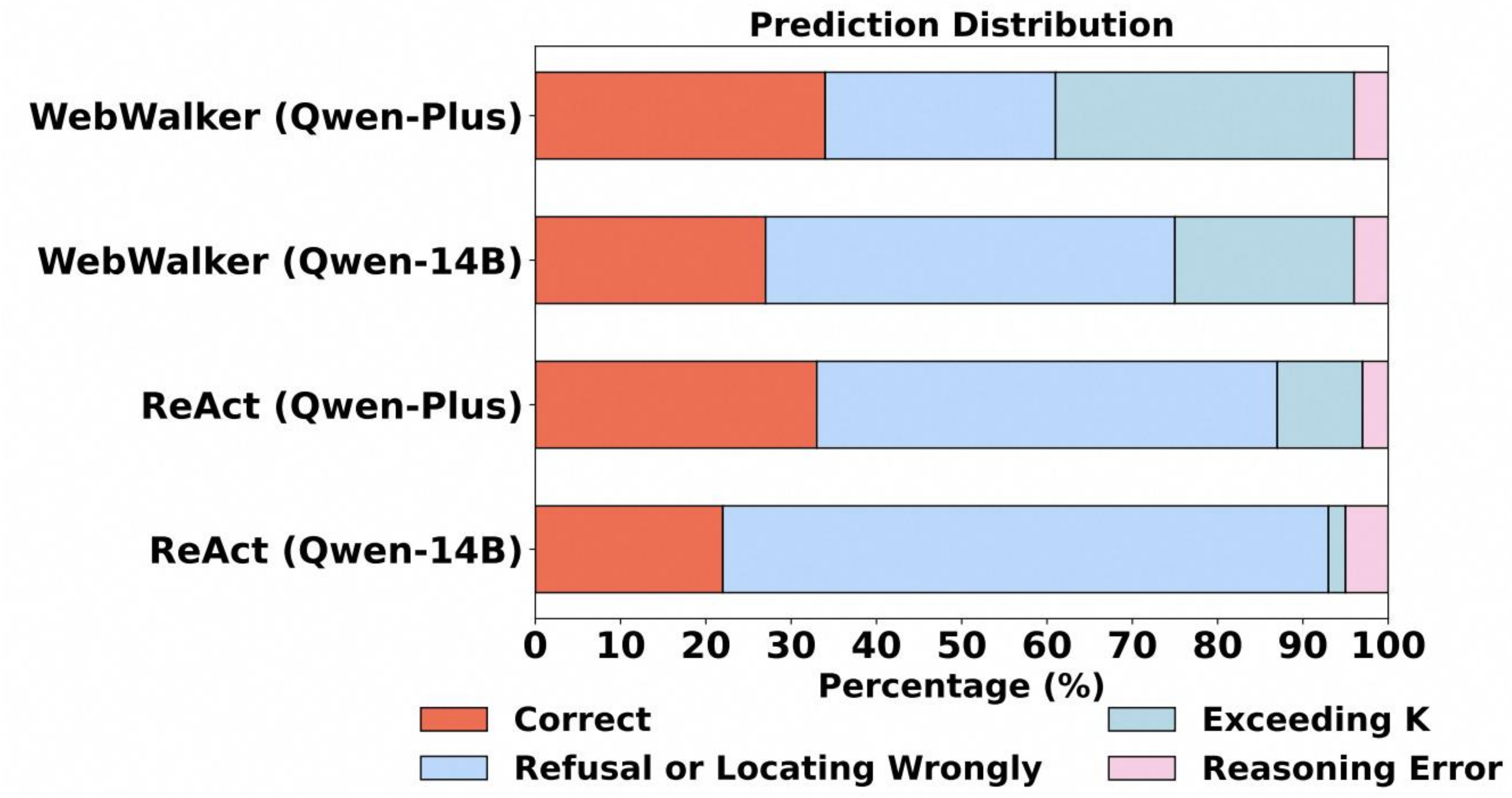
Acc. and Action Count **distribution.**



Performance across domains and languages.

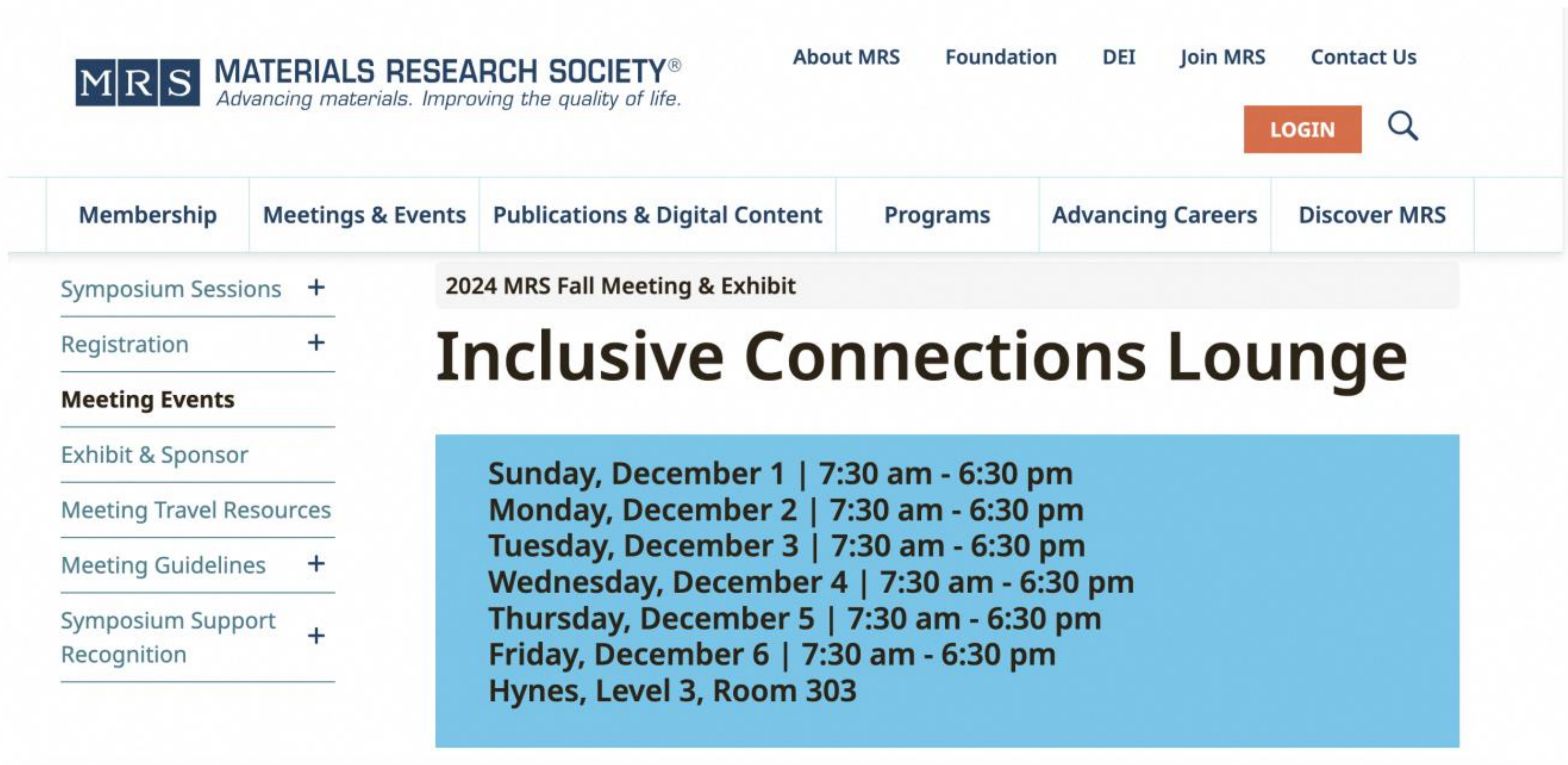
RESULTS AND DISCUSSION

Results on Agents and RAG systems



Predication distribution of WebWalker and ReAct.

A case requiring **reasoning capability**.

Root Url	`https://www.mrs.org/`
Question	How many hours in total would a person spend if they attended the **Inclusive Connections Lounge** activities from December 1 to 6, 2024, at the MRS Fall Meeting?
Answer	66 hours
Source Website	`https://www.mrs.org/meetings-events/annual-meetings/2024-mrs-fall-meeting/meeting-events/broadening-participation/inclusive-connections-lounge`
Website Information	

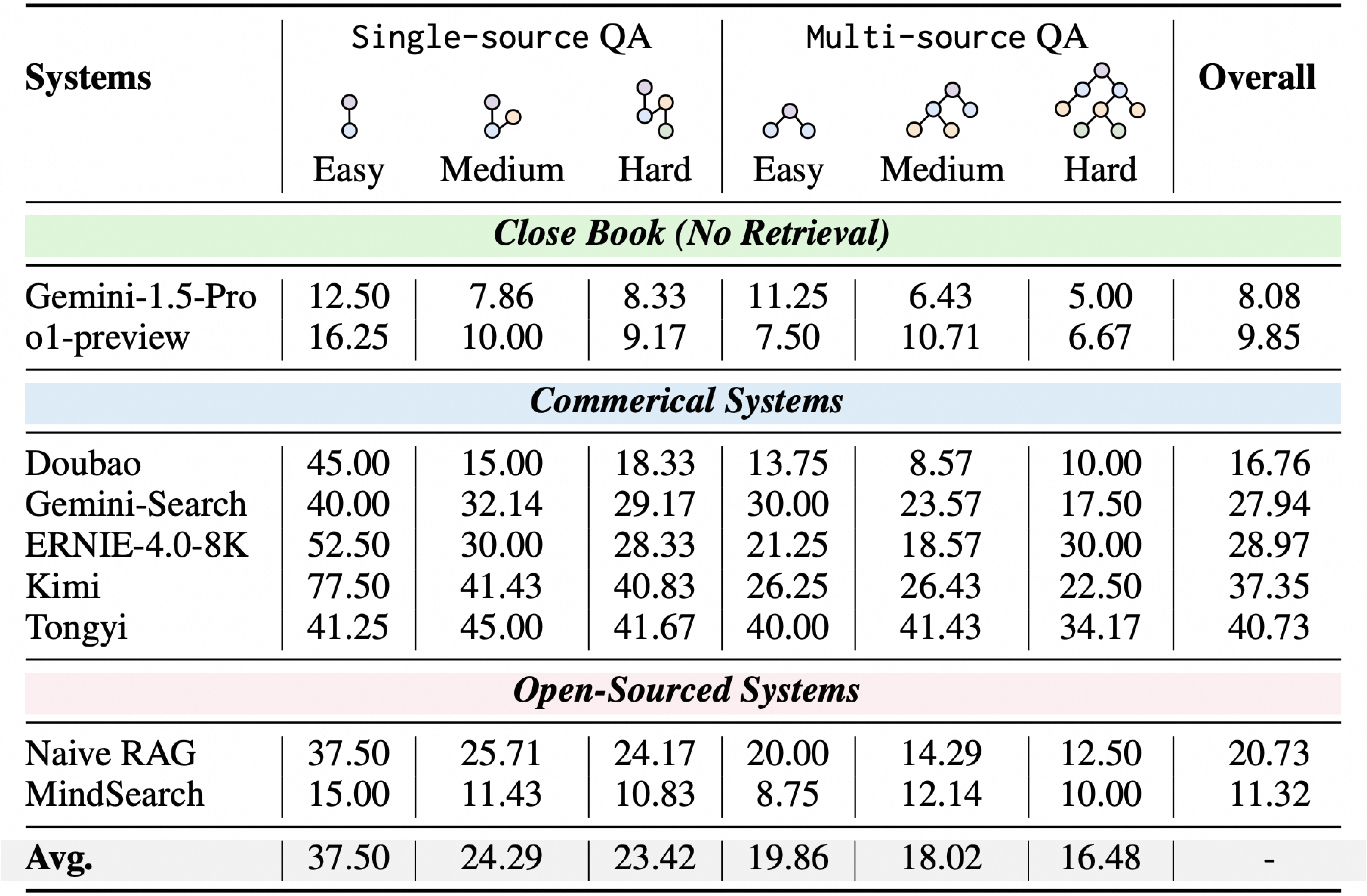
2024 MRS Fall Meeting & Exhibit

Inclusive Connections Lounge

Sunday, December 1 | 7:30 am - 6:30 pm
Monday, December 2 | 7:30 am - 6:30 pm
Tuesday, December 3 | 7:30 am - 6:30 pm
Wednesday, December 4 | 7:30 am - 6:30 pm
Thursday, December 5 | 7:30 am - 6:30 pm
Friday, December 6 | 7:30 am - 6:30 pm
Hynes, Level 3, Room 303

RESULTS AND DISCUSSION

Results on Agents and RAG systems



Systems	Single-source QA			Multi-source QA			Overall
	Easy	Medium	Hard	Easy	Medium	Hard	
Close Book (No Retrieval)							
Gemini-1.5-Pro	12.50	7.86	8.33	11.25	6.43	5.00	8.08
o1-preview	16.25	10.00	9.17	7.50	10.71	6.67	9.85
Commerical 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	-

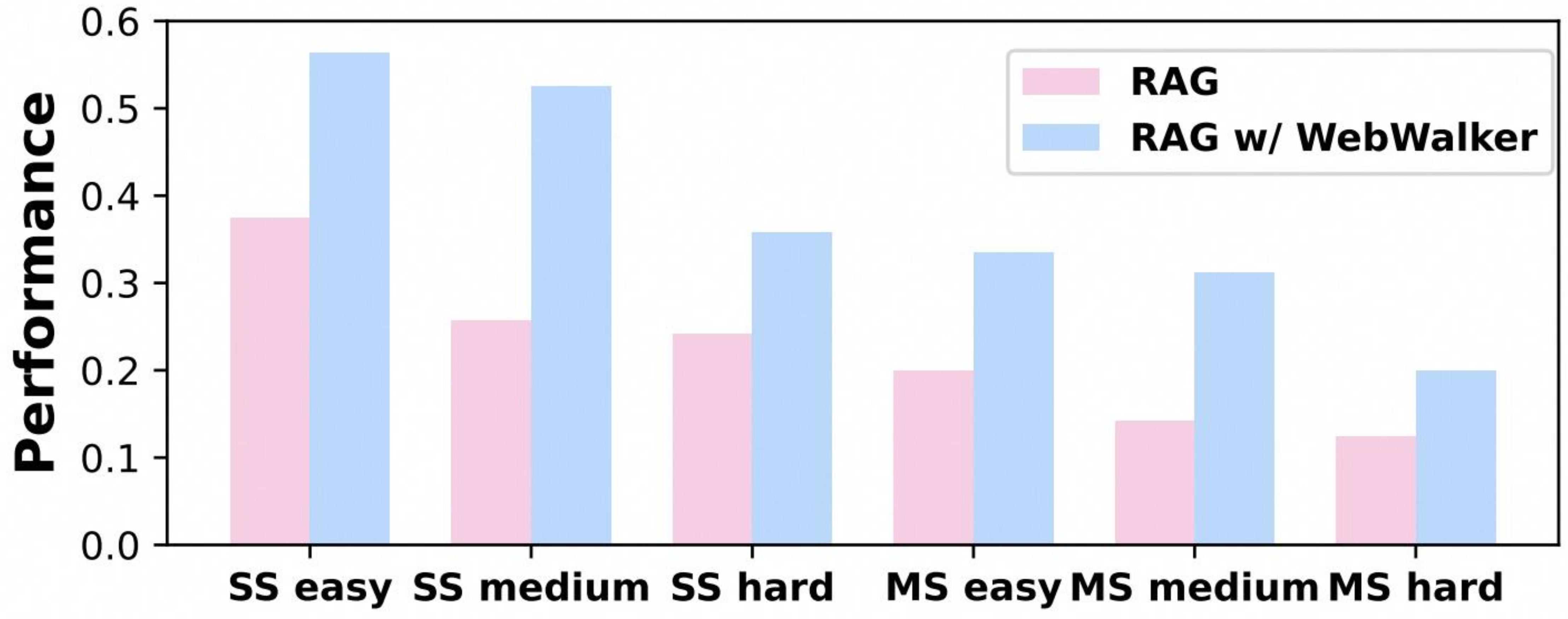
Main results of RAG systems on WebWalkerQA.

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

RESULTS AND DISCUSSION

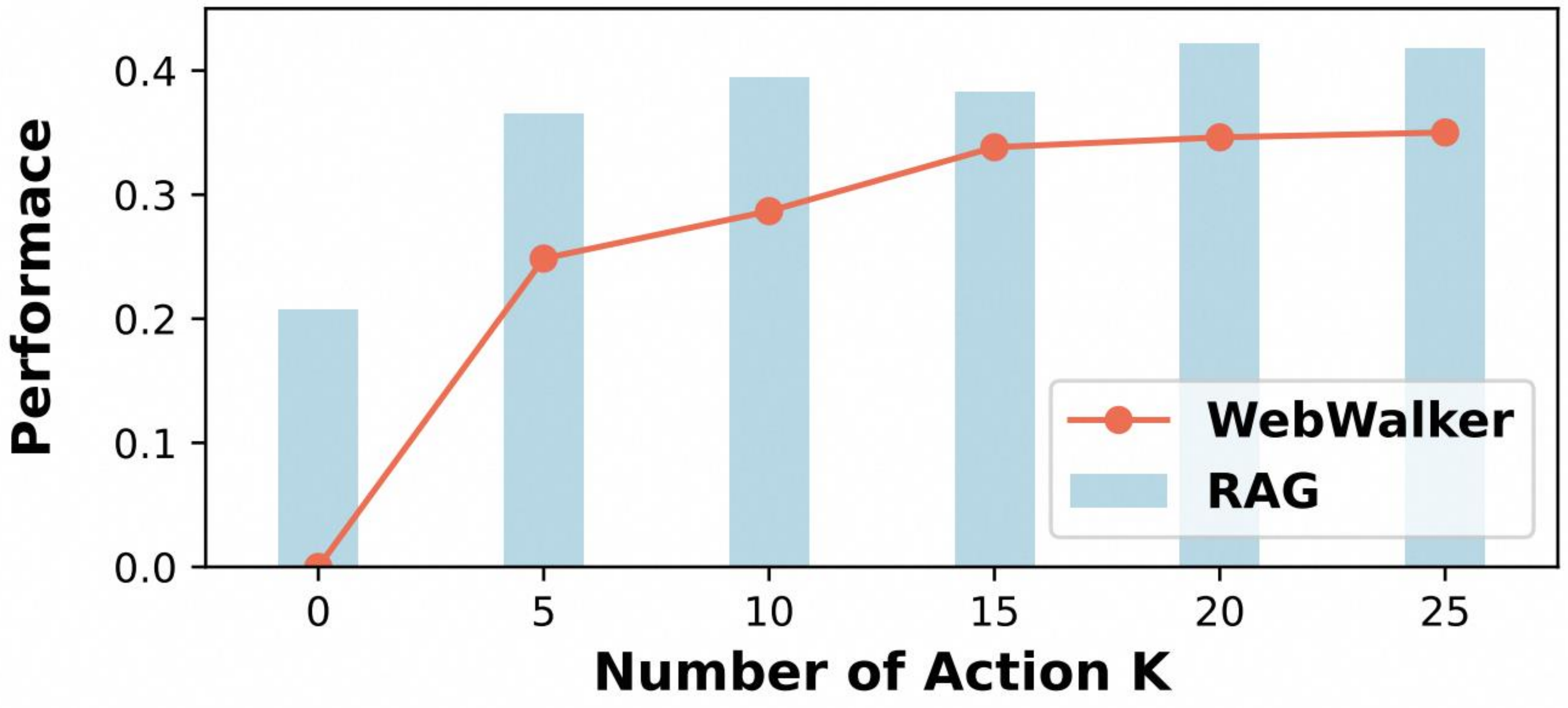
Results on Agents and RAG systems

WebWalker Combined with RAG System



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

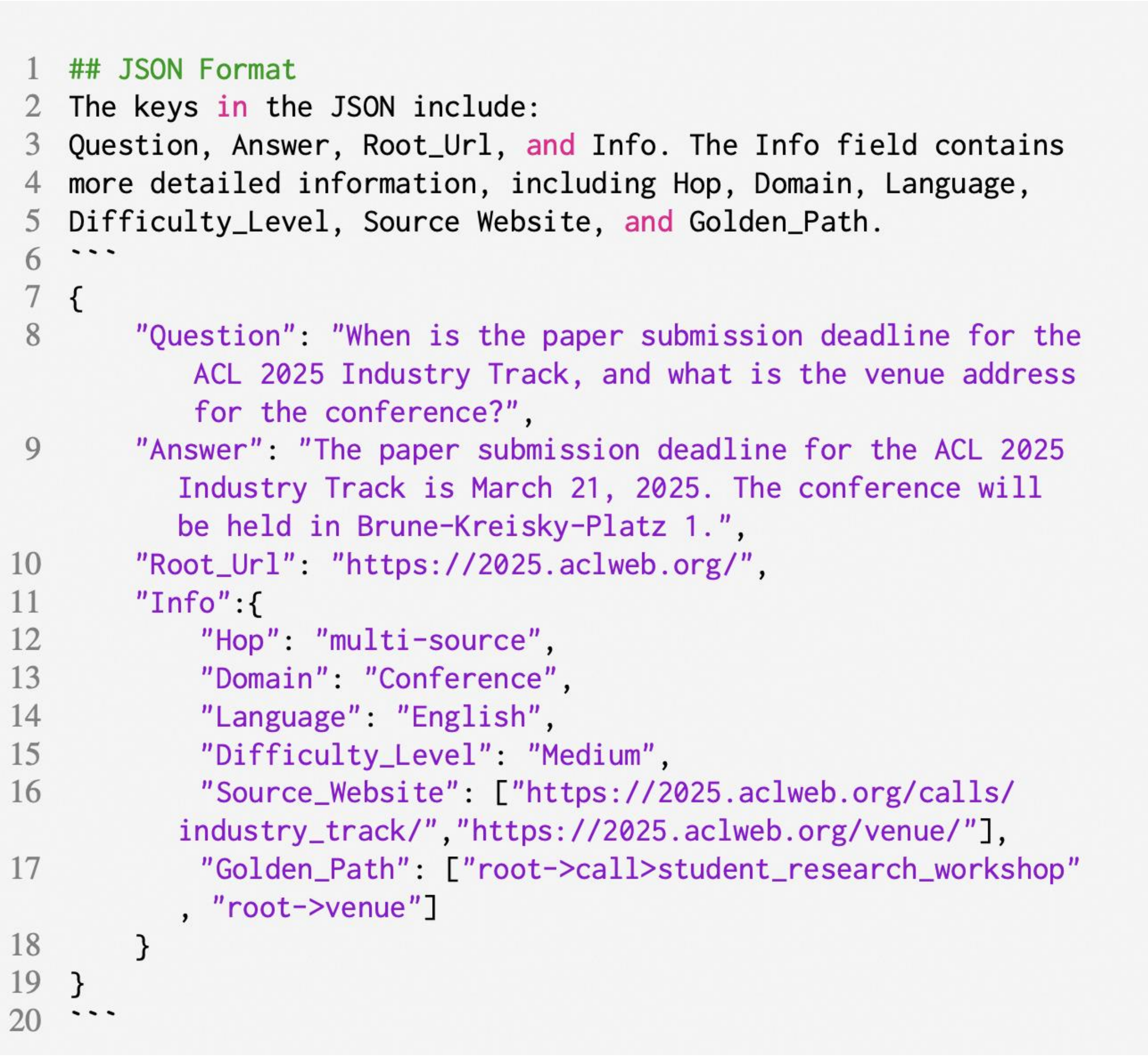
Scaling Up on Action Count K



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

CONCLUSION

Limitations and Future Works



~~~
## JSON Format
The keys in the JSON include:
Question, Answer, Root_Url, and Info. The Info field contains
more detailed information, including Hop, Domain, Language,
Difficulty_Level, Source Website, and Golden_Path.
```
{
    "Question": "When is the paper submission deadline for the
        ACL 2025 Industry Track, and what is the venue address
        for the conference?",
    "Answer": "The paper submission deadline for the ACL 2025
       Industry Track is March 21, 2025. The conference will
       be held in Brune-Kreisky-Platz 1.",
    "Root_Url": "https://2025.aclweb.org/",
    "Info":{
        "Hop": "multi-source",
        "Domain": "Conference",
        "Language": "English",
        "Difficulty_Level": "Medium",
        "Source_Website": ["https://2025.aclweb.org/calls/
       industry_track/","https://2025.aclweb.org/venue/"],
        "Golden_Path": ["root->call>student_research_workshop"
       , "root->venue"]
    }
}
```
~~~

Takeaways:

- A QA-format Web Traversal dataset.

  🗄️ Datasets

- A multi-agent framework.

- Insights of information seeking through horizontal and vertical integration.

- We open-source on **https://github.com/Alibaba-NLP/WebWalker.**

CONCLUSION

Limitations and Future Works

Dataset Size: 680 -> 14k silver data

Multimodal Environment: screenshots or GUI

Agent Tuning: RL for Web agents (more browser actions)

Better Integration with RAG Systems: deep research

通义

致力于实现类人智慧的通用智能_

通义