R-WoM: Retrieval-augmented World Model for Computer-use Agents

Kai Mei[1,2], Jiang Guo[2*], Shuaichen Chang[2], Mingwen Dong[2],
Dongkyu Lee[2], Xing Niu[2], Jiarong Jiang[2]

[1] RUTGERS [2] aws

Presenter: Kai Mei

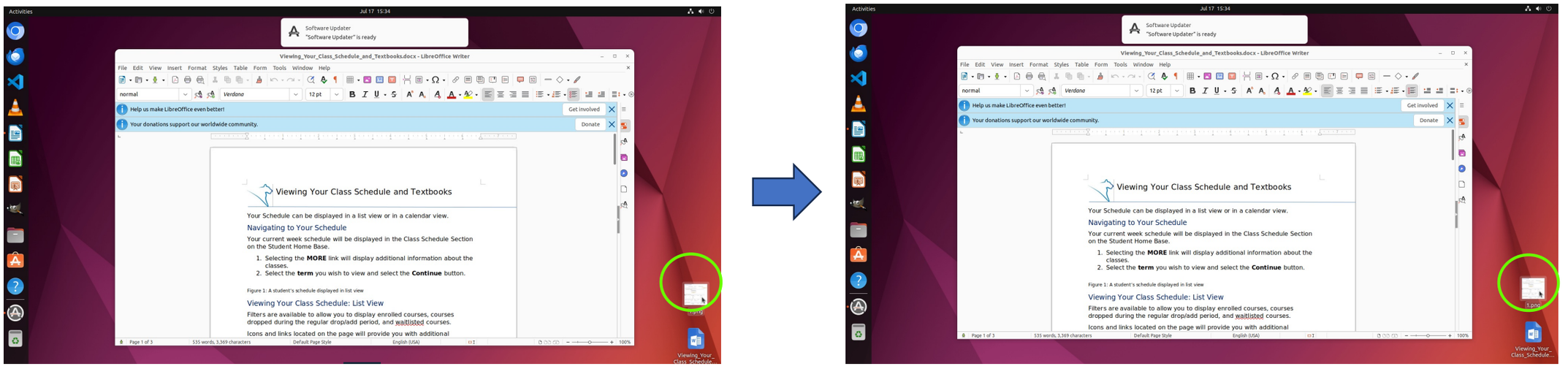
kai.mei@rutgers.edu

*: Corresponding author. Work done when Kai was an intern at AWS Agentic AI.

Knowledge Gap Between LLMs and Envs.

Task: Copy the screenshot 1.png from the desktop to where my cursor is located

From the LLM's perspective

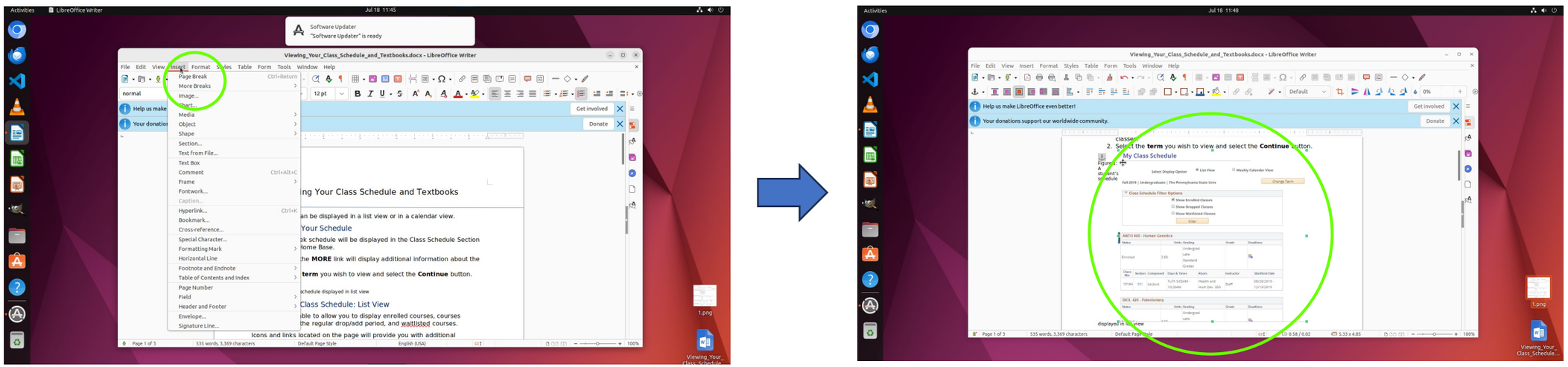


Failed as it loses the cursor position

Knowledge Gap Between LLMs and Envs.

Task: Copy the screenshot 1.png from the desktop to where my cursor is located

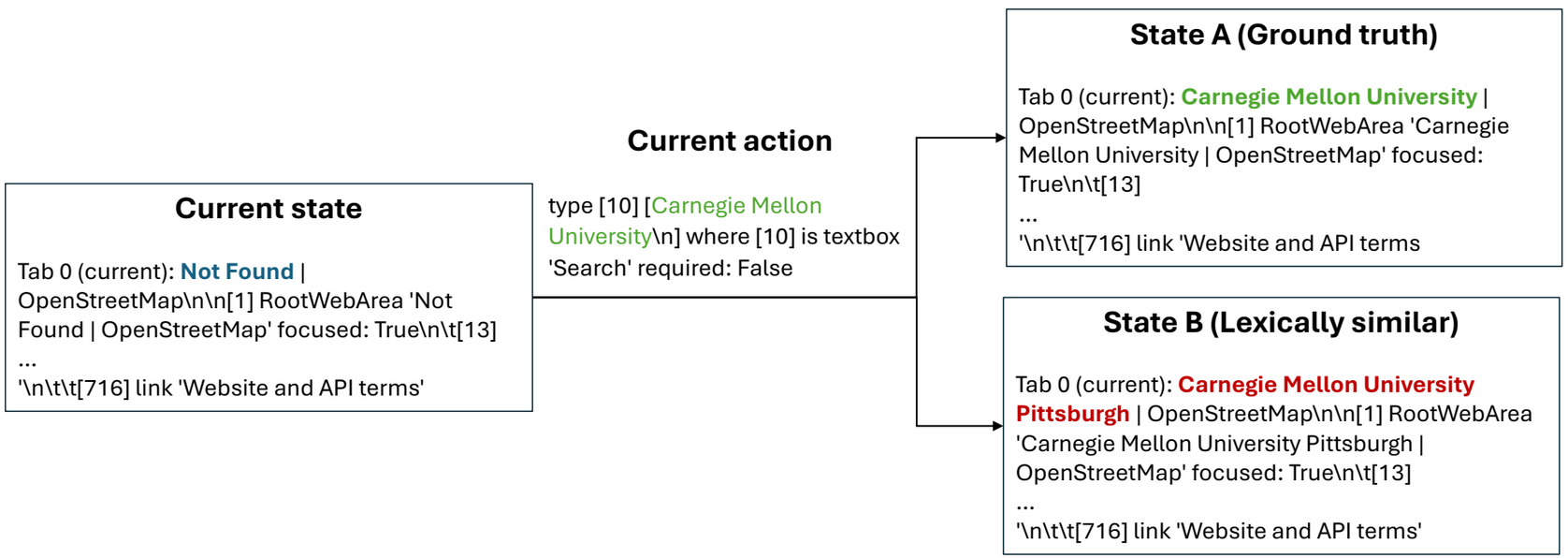
From the Environment's perspective



Succeed by inserting images

Knowledge Probing of LLMs

Task 1: Next State Identification



Knowledge Probing of LLMs

Task 2: Full-procedure Planning Alignment

Task: Tell me the full name, gitlab account name, location and email address of the contributor who has the most commits to branch php52

Generated Plan

1. Initial navigation steps:
- Look for and click on "Branches" in the left sidebar menu
- This should take us to a branches overview page

2. On the branches page:
- Search for "php52" branch in the branch list
- Note: If the branch is not visible, we may need to:
a. Use the search functionality within branches
b. Check if we have proper access to view all branches
c. Verify if the branch exists

3. If we find the php52 branch:
- Click on the branch name to view its details
- Look for a "Contributors" or similar tab/section

4. In the contributors section:
- Sort contributors by number of commits (if such sorting is available)
- Identify the contributor with the most commits

5. For the identified top contributor:
- Click on their profile link/name to view their details
- Look for their full name, location, and email information

Whether aligned?

Tutorials

View contributor analytics

To view contributor analytics for a project:

1. On the left sidebar, select **Search or go to** and find your project.
2. Select **Analyze > Contributor analytics**.
...

View project commit history

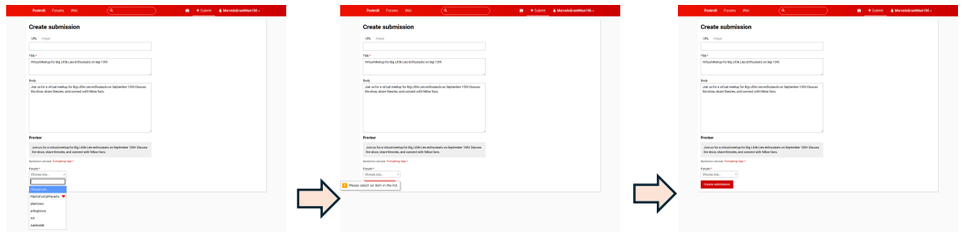
To view a list of commits made by project members per day:

1. On the left sidebar, select **Search or go to** and find your project.
1. Select **Analyze > Contributor analytics**.
1. Select **History**.
1. From the **Branches** (**main**) dropdown list, select the branch you want to view commits for.
1. To view the number of commits made by the members on a specific day, hover over the line chart.
Optional. Filter the results.

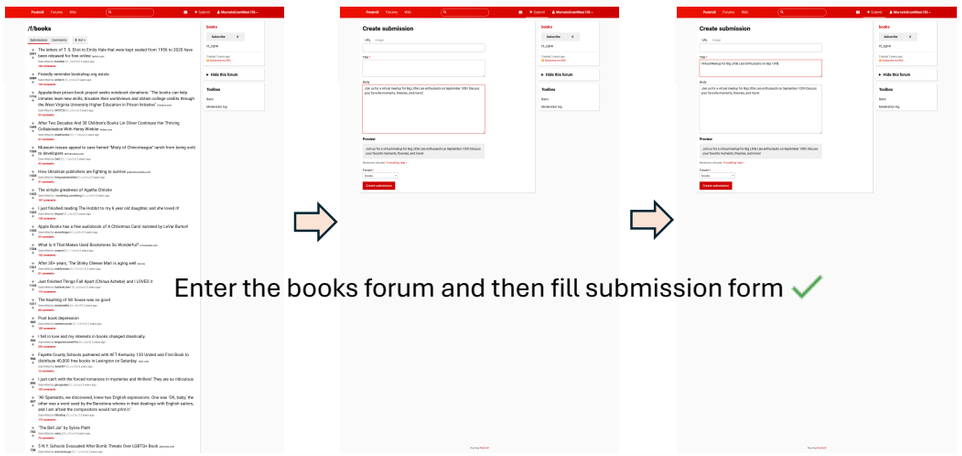
- To filter by author, from the **Author** dropdown list, select the user whose commits you want to view.
- To filter by commit message, in the text box, enter your search criteria.

Knowledge Probing of LLMs

Task 3: Milestone Transition Recognition



Get stuck in selecting books as the forum type to create submission ❌



Enter the books forum and then fill submission form ✓

Knowledge Probing of LLMs

Table 1: Probing results across three tasks: next-state identification, full-procedure planning alignment, and milestone transition recognition. All values are percentages.

Model	Next-state identification (by lexical similarity)				Full-procedure planning alignment		Milestone transition recognition
	[0, 0.8)	[0.8, 0.9)	[0.9, 1]	Overall	w/o retrieval	w/ retrieval	Accuracy
Qwen-2.5-VL-72B	61.1	84.8	77.6	77.0	50.0	90.0	83.7
Claude-3.5-Sonnet	72.2	84.8	81.6	81.0	55.0	85.0	85.7
Claude-3.7-Sonnet	88.9	87.9	83.7	86.0	65.0	95.0	86.7

> 75 < 65 > 80

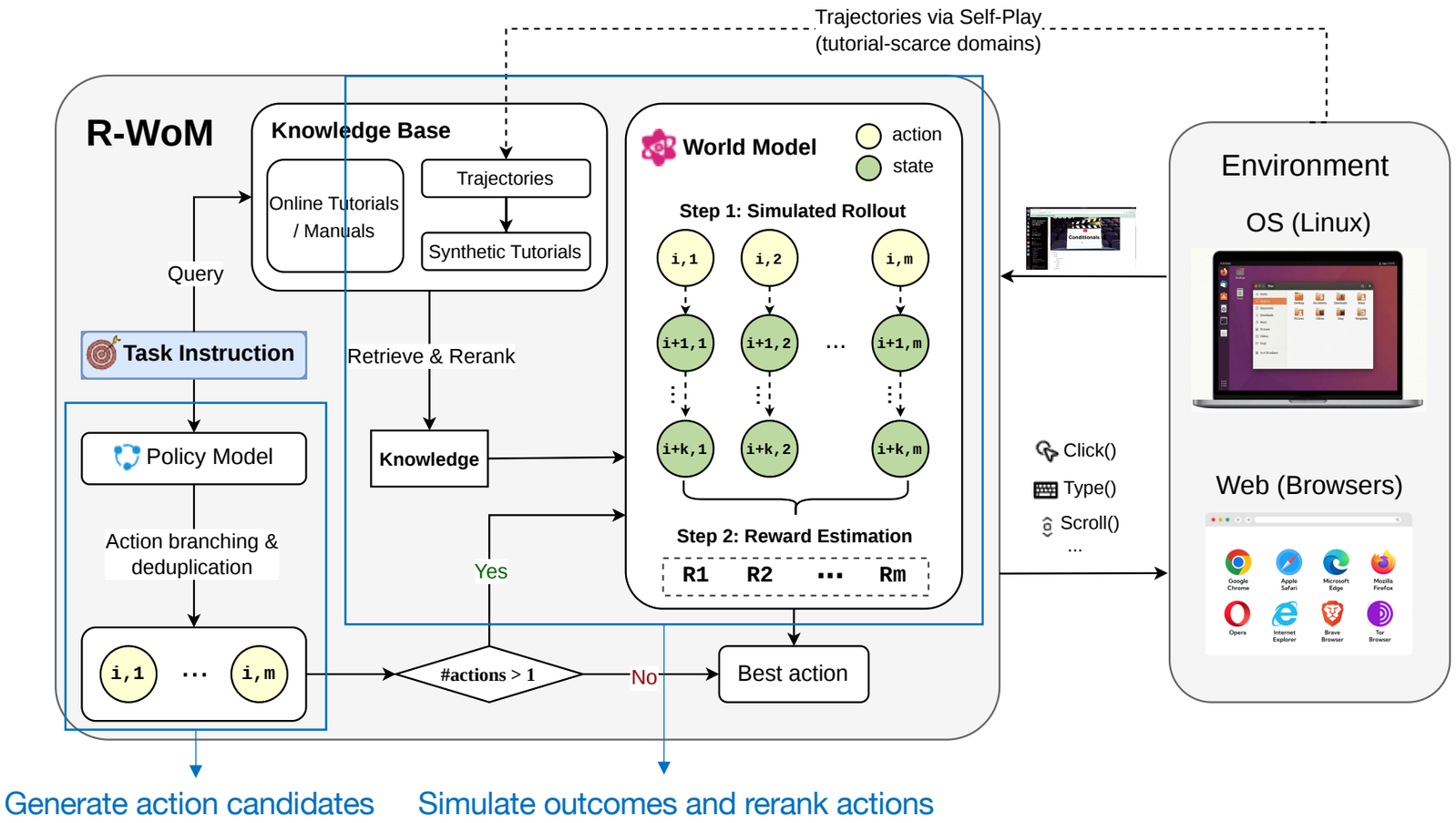
Takeaways: LLMs can be good at understanding short-term outcomes, but these strengths do not extend to long-horizon planning to be aligned with the environment dynamics.

Grounding Considerations

- **Thought 1**: How can we extract more accurate and reliable knowledge from environment dynamics?
  - Retrieval augmentation (e.g., query rewriting, reranking)

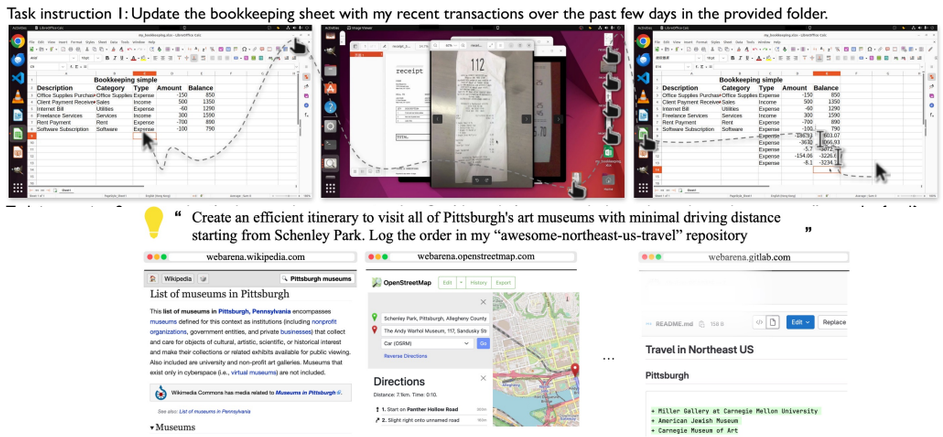
- **Thought 2**: How can we better leverage this structured knowledge to guide LLMs at inference time?
  - Explicit reasoning over the potential outcomes

Grounding Framework (R-WoM)



How Grounding Performs?

- Benchmarks
  - OSWorld
  - WebArena
- Model
  - Qwen2.5-72B, Claude-Sonnet (3.5, 3.7)
- Baselines
  - Vanilla, RAG, WebDreamer



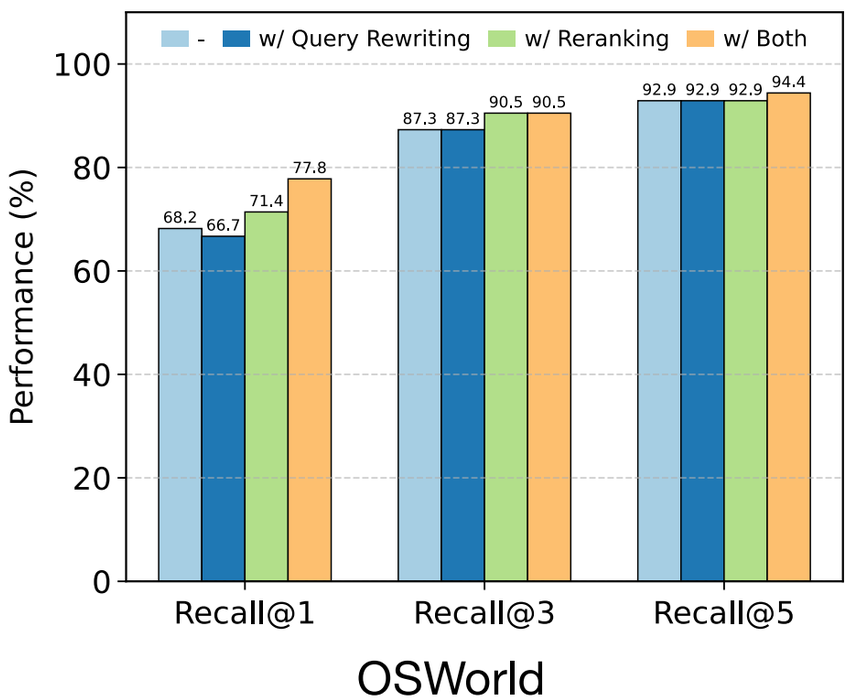
How Grounding Performs?

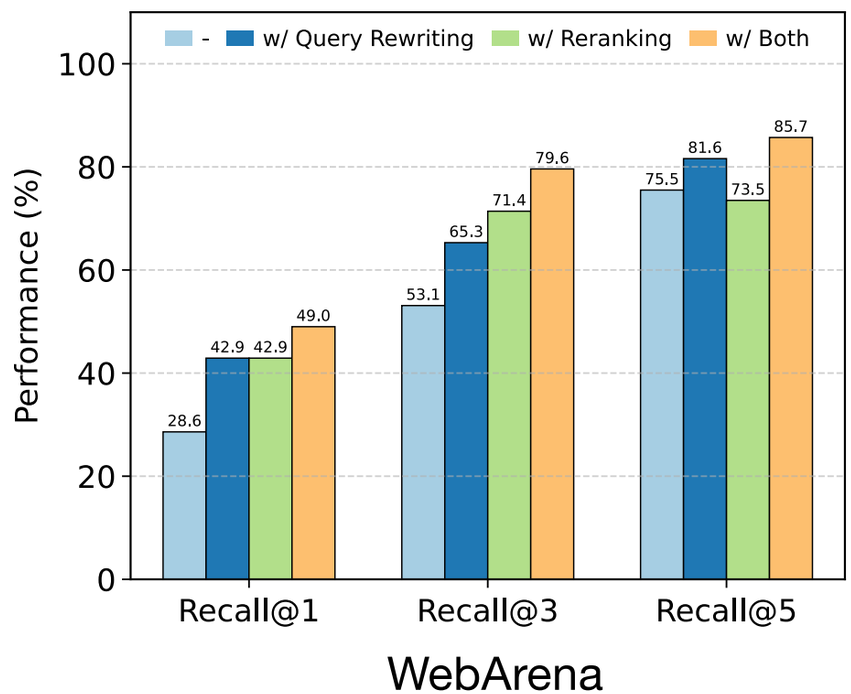
Table 2: End-to-end performance on OSWorld and WebArena across three runs. Best in **bold**; second-best underlined. ↑· denotes this relative improvement over the second-best baseline

Model	Method	OSWorld (Xie et al., 2024)	WebArena (Zhou et al., 2023)
Qwen-2.5-VL-72B	Vanilla	26.36 ± 2.32	21.84 ± 0.42
	RAG	30.84 ± 1.07	22.42 ± 0.42
	WebDreamer	28.37 ± 2.01	24.50 ± 0.84
	R-WoM	**37.48 ± 2.29** ↑21.5%	**28.49 ± 0.43** ↑16.3%
Claude-3.5-Sonnet	Vanilla	22.43 ± 2.25	27.74 ± 0.43
	RAG	22.19 ± 0.92	30.70 ± 0.41
	WebDreamer	23.48 ± 2.14	29.82 ± 0.41
	R-WoM	**26.01 ± 0.44** ↑10.8%	**33.15 ± 0.01** ↑8.0%
Claude-3.7-Sonnet	Vanilla	28.47 ± 2.27	28.92 ± 0.41
	RAG	27.76 ± 0.75	32.75 ± 0.72
	WebDreamer	31.24 ± 2.88	31.86 ± 0.01
	R-WoM	**38.54 ± 1.92** ↑23.4%	**34.58 ± 1.10** ↑5.6%

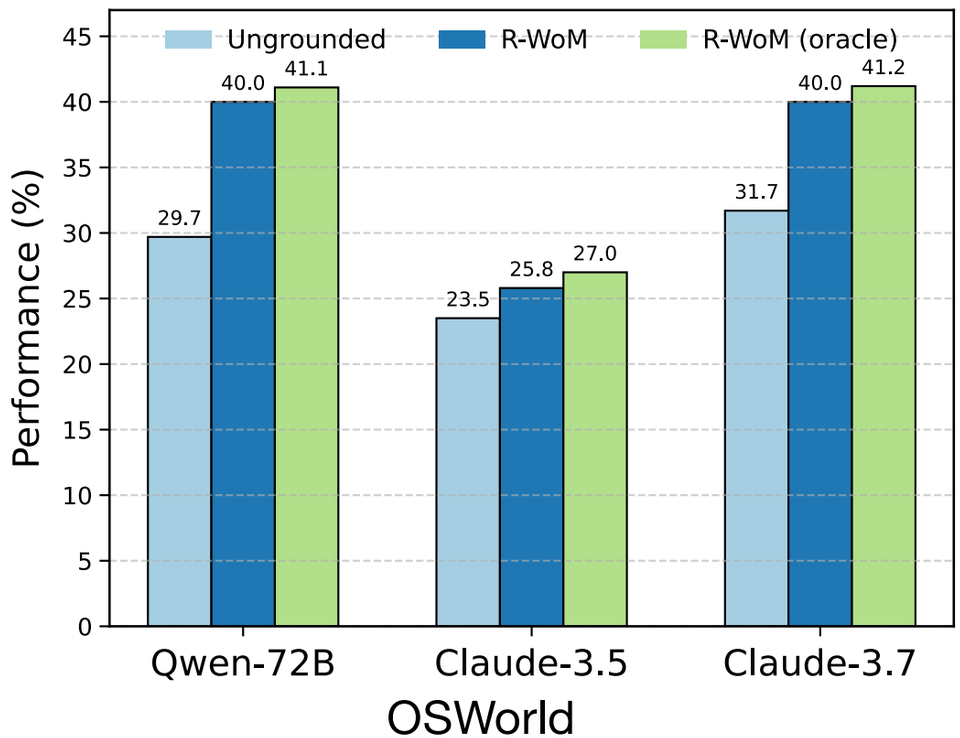
Ablations on the Grounding Quality

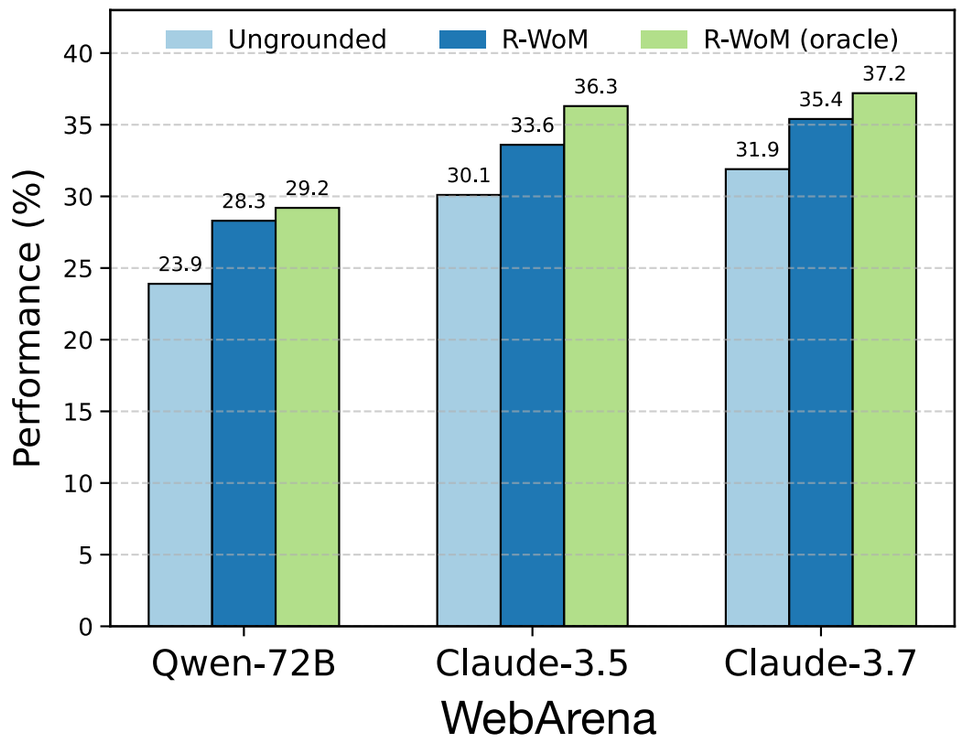
How reliable is the knowledge retrieved for grounding?





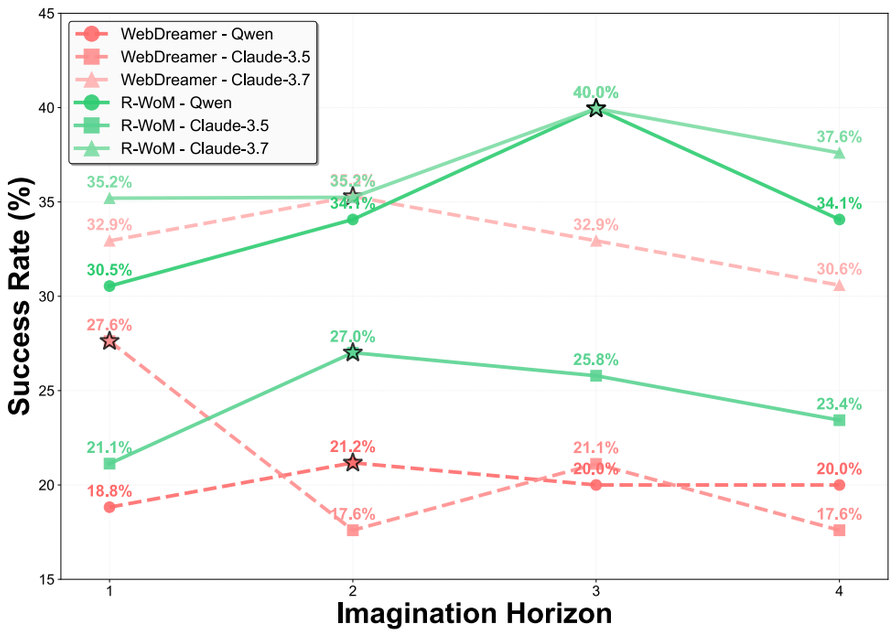
Ablations on the Grounding Quality



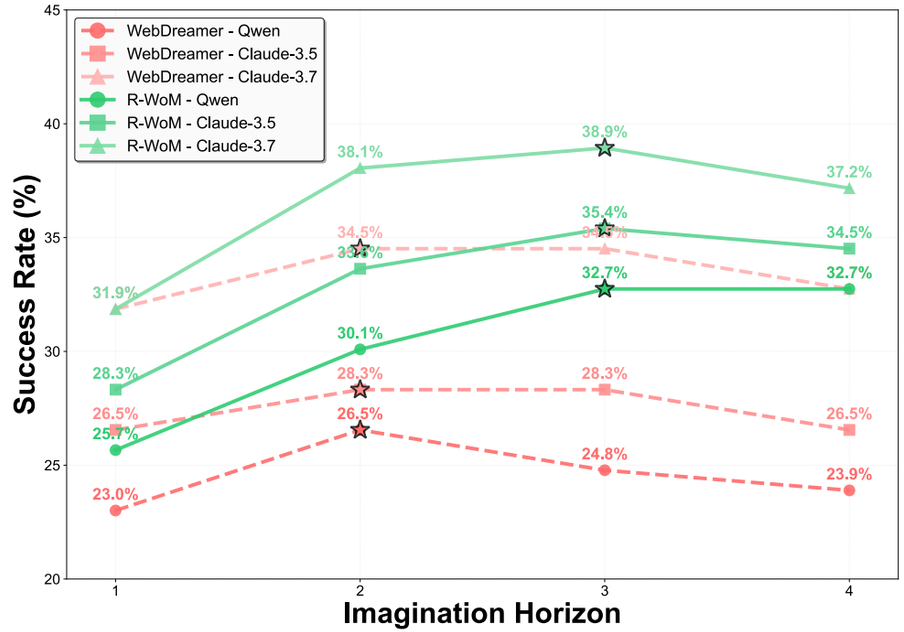


Takeaways: Grounding quality is closely related to knowledge relevance

The Potential of Grounding



OSWorld



WebArena

Takeaways: Grounding makes it easier for LLMs to be aware of the action consequences, but errors still accumulate in predicting the far future

Future Work

- Knowledge-transfer
  - Explore whether the environment-specific knowledge can be transferred to completely undocumented, proprietary environments with no prior grounding data.

- Meta-Learning for Knowledge-awareness
  - Developing an architecture where the agent can be aware of the knowledge gap and deal with potential knowledge conflicts

Thank you!