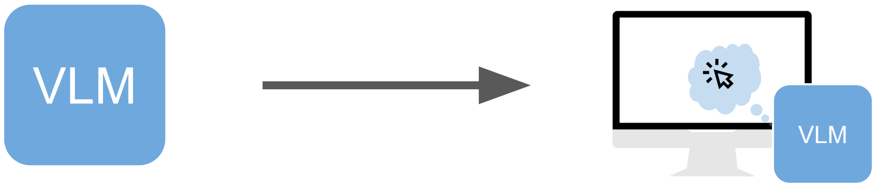
AI Agents Tutorial

From next token prediction to digital automation



Xiao Yu, Zhou Yu
Columbia University & Articulate AI

First part of the slides are adapted from [Shunyu Yao's PhD Thesis Defense]

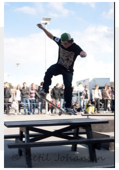
Interacting with the Digital World using VLMs

Time

2020-2022

VQA Tasks

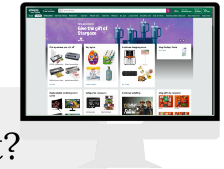
Q: What is he doing?



He is performing a skateboard trick...

2023+

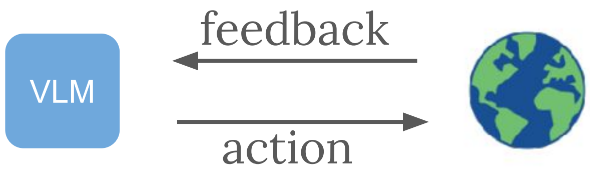
Computer Tasks



Can you help me clear my shopping cart?

click button [shopping cart] ....

Interacting with the Digital World using VLMs



Agent	Environment
(V)LM Chatbots	Interact with human / answer questions
Robotics Agents	Interact with a physical world
Visual Language Agents	**Interact with digital devices (e.g., your phone)**

Building "Agents" in other domains

Challenge 1: need accessible methods to build general agents

Intensive to build

(Even for experts)

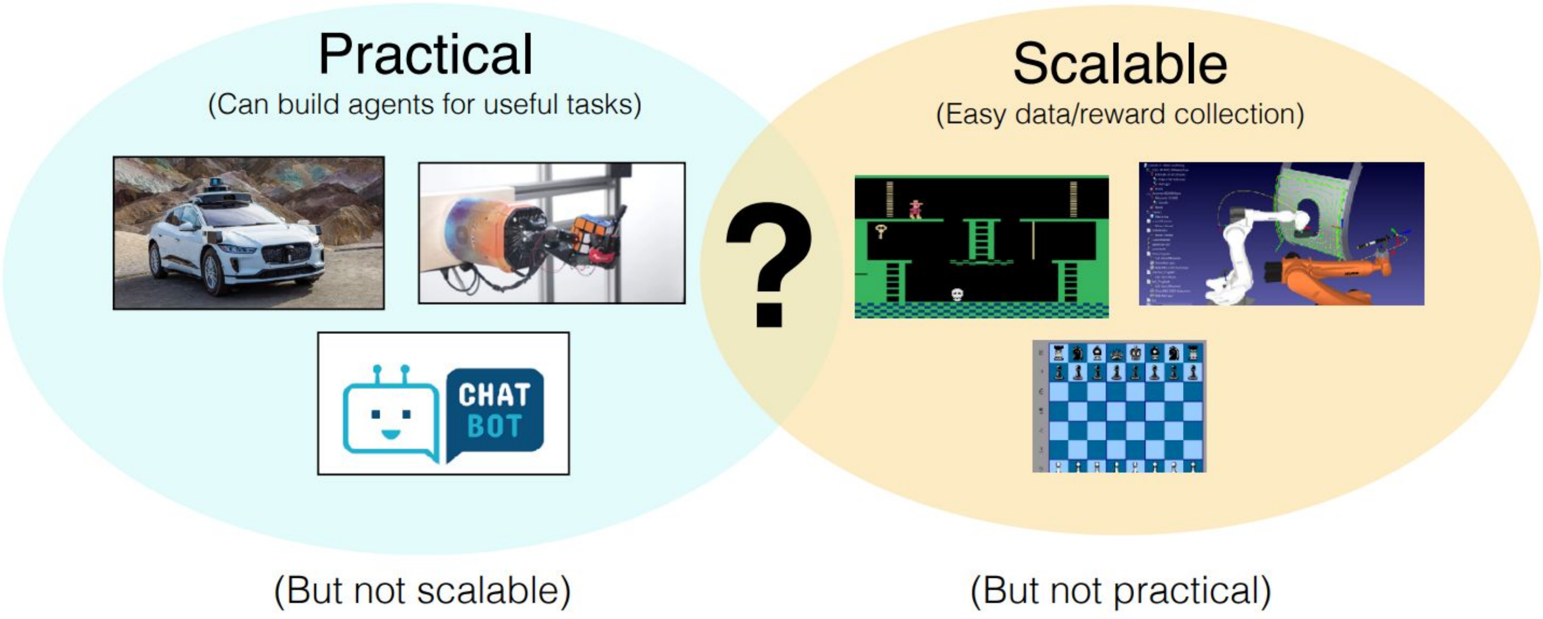
Takes millions of lines of rules (by domain experts)

Takes millions of training iterations (by RL experts)

Hard to generalize

Building "Agents" in other domains

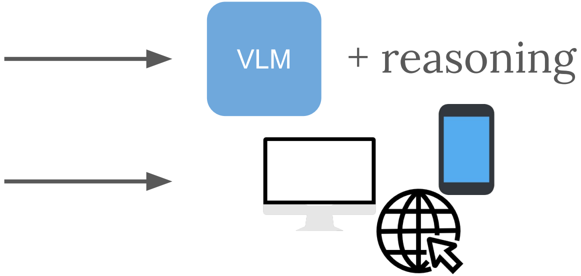
Challenge 2: need scalable benchmarks for practical tasks

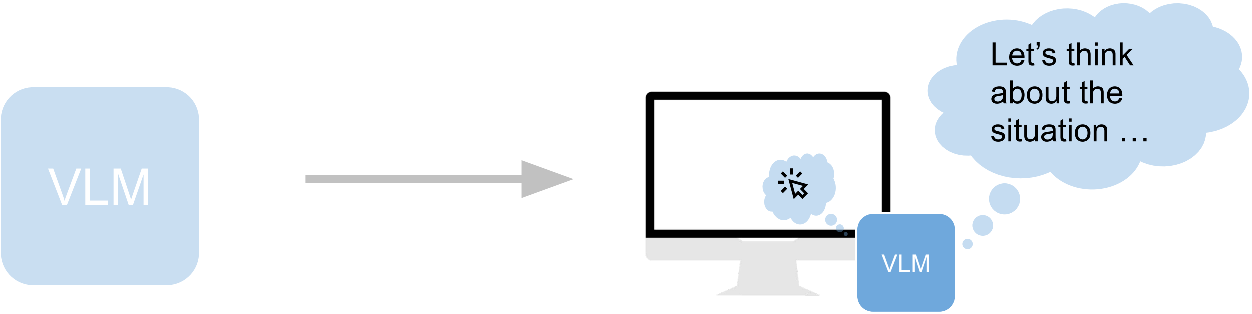


Interacting with the Digital World using VLMs

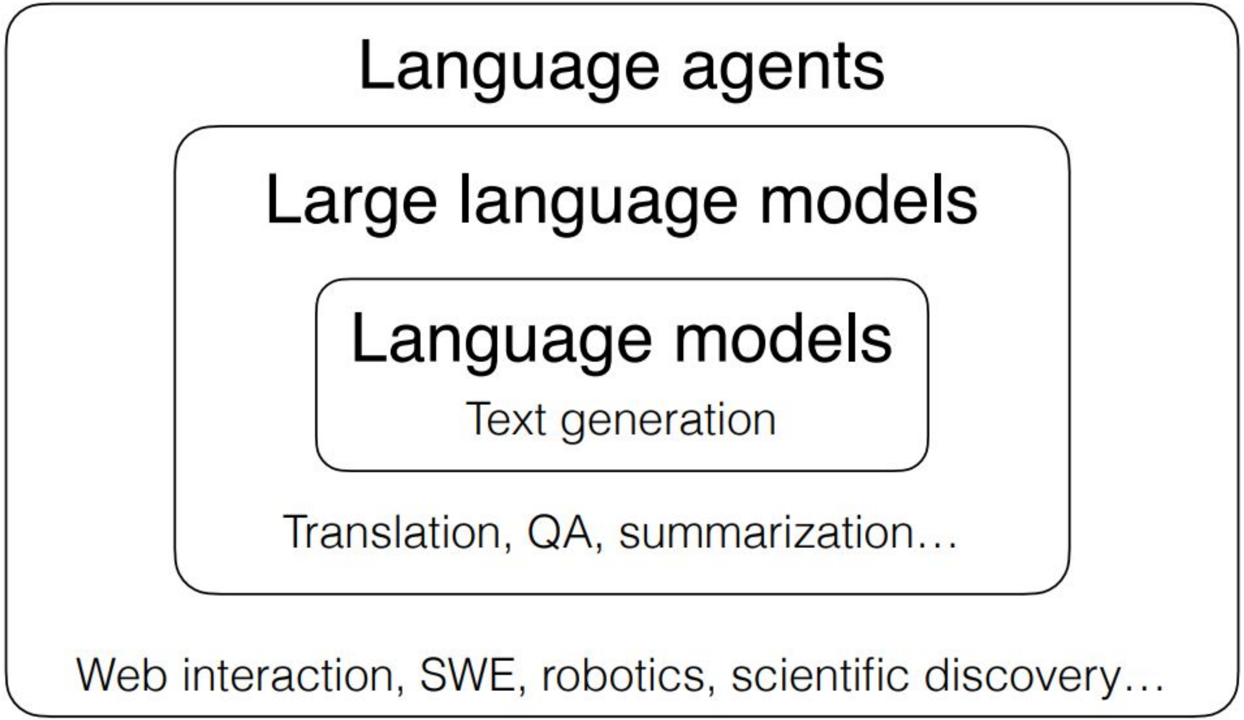
Challenge 1: … build general agents

Challenge 2: … scalable benchmarks for practical tasks





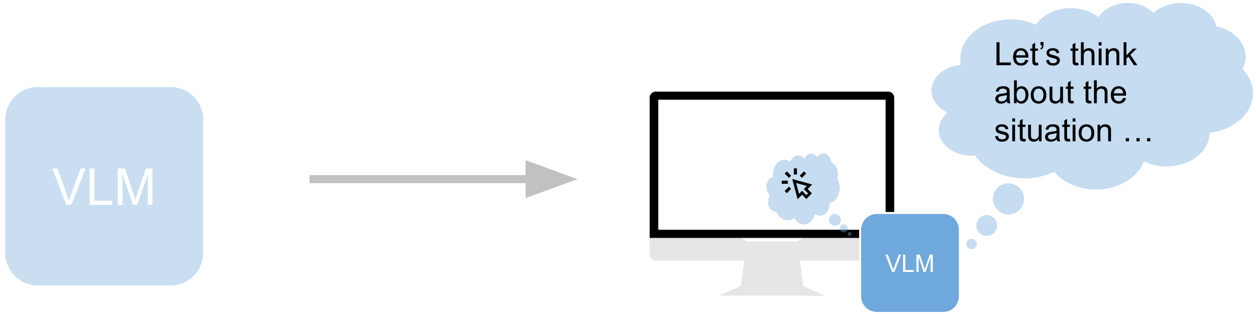
Interacting with the Digital World using VLMs



Today's Lecture

Part 1. Benchmarking VLM's performance on digital tasks

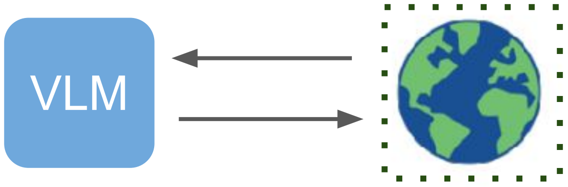
Part 2: Building (visual) language agents for device control



1

Benchmarking VLM's performance on digital tasks

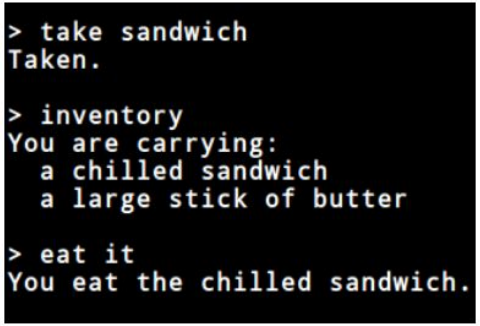
Realistic environments based on computer/web/android devices



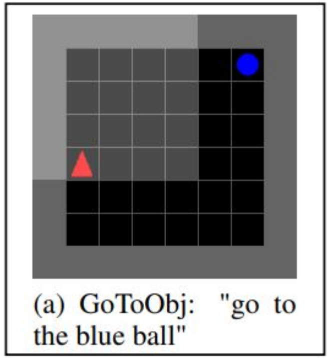
Agent Benchmarks with Device Control

Related benchmarks prior to 2020

MiniWoB
(Shi et al., 2017)



TextWorld
(Côté et al., 2019)

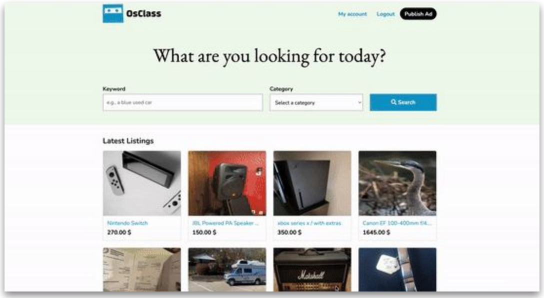


BabyAI
(Chevalier-Boisvert et al., 2019)

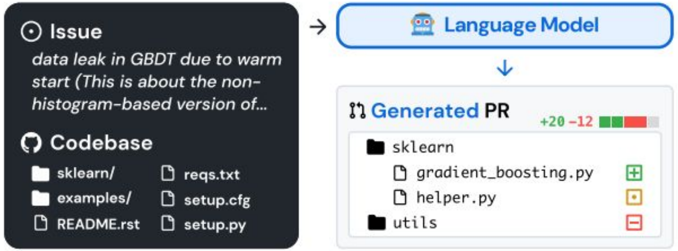
- Simulation environment
- Synthetic text (if any)
- Small action space
- Short-horizon tasks

Agent Benchmarks with Device Control

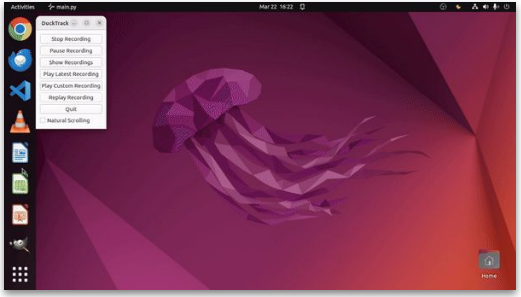
Recent benchmarks (non-exhaustive)



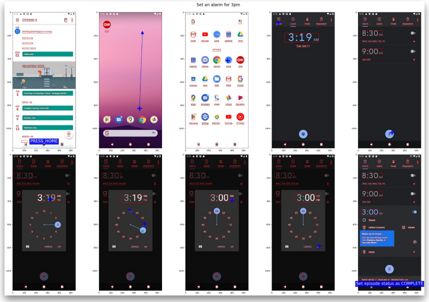
WebArena/VisualWebArena



SWE-Bench



OSWorld



Android in the Wild

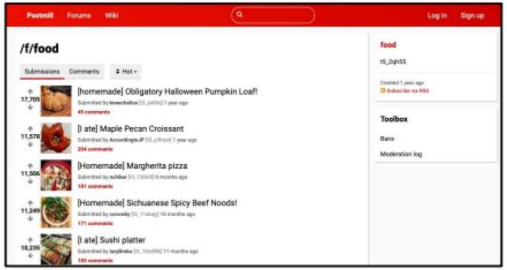
Agent Benchmarks with Device Control

WebArena/VisualWebArena

Goal: Evaluate VLM’s ability to navigate on the web

Environment: Locally hosted websites cloned from real services (e.g., reddit)

Task Input: instruction + initial web screenshot



Original Webpage

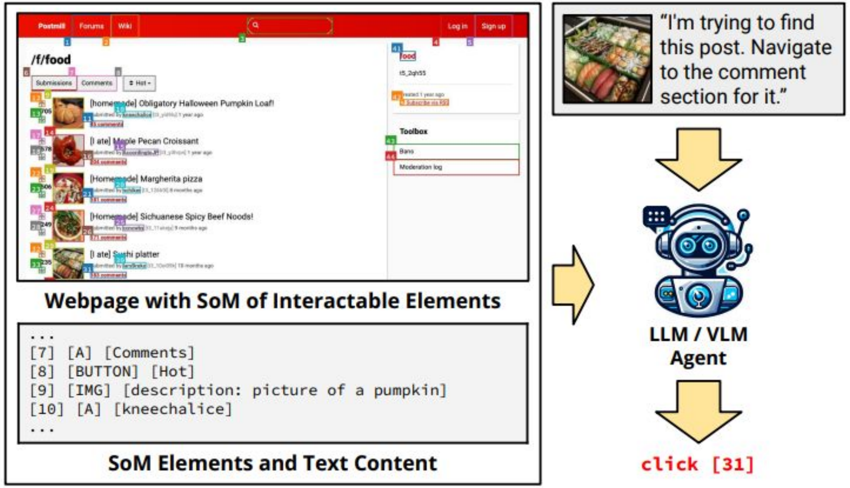
Evaluation: pre-defined scripts to check final state of the website

Zhou, Shuyan, et al. "Webarena: A realistic web environment for building autonomous agents." arXiv preprint arXiv:2307.13854 (2023).
Koh, Jing Yu, et al. "Visualwebarena: Evaluating multimodal agents on realistic visual web tasks." arXiv preprint arXiv:2401.13649 (2024).

Agent Benchmarks with Device Control

WebArena/VisualWebArena

Baselines: convert clickable buttons to IDs, and prompt an VLM to output which ID to click next



Zhou, Shuyan, et al. "Webarena: A realistic web environment for building autonomous agents." arXiv preprint arXiv:2307.13854 (2023).
Koh, Jing Yu, et al. "Visualwebarena: Evaluating multimodal agents on realistic visual web tasks." arXiv preprint arXiv:2401.13649 (2024).

Agent Benchmarks with Device Control

WebArena/VisualWebArena

Results: current state-of-the-art VLMs significantly lag behind human

Model Type	LLM Backbone	Visual Backbone	Inputs	Success Rate (↑)			
				Classifieds	Reddit	Shopping	Overall
Text-only	LLaMA-2-70B	-	Acc. Tree	0.43%	1.43%	1.29%	1.10%
	Mixtral-8x7B			1.71%	2.86%	1.29%	1.76%
	Gemini-Pro			0.85%	0.95%	3.43%	2.20%
	GPT-3.5			0.43%	0.95%	3.65%	2.20%
	GPT-4			5.56%	4.76%	9.23%	7.25%
Caption-augmented	LLaMA-2-70B	BLIP-2-T5XL	Acc. Tree + Caps	0.00%	0.95%	0.86%	0.66%
	Mixtral-8x7B	BLIP-2-T5XL		1.28%	0.48%	2.79%	1.87%
	GPT-3.5	LLaVA-7B		1.28%	1.43%	4.08%	2.75%
	GPT-3.5	BLIP-2-T5XL		0.85%	1.43%	4.72%	2.97%
	Gemini-Pro	BLIP-2-T5XL		1.71%	1.43%	6.01%	3.85%
	GPT-4	BLIP-2-T5XL		8.55%	8.57%	16.74%	12.75%
Multimodal	IDEFICS-80B-Instruct		Image + Caps + Acc. Tree	0.43%	0.95%	0.86%	0.77%
	CogVLM			0.00%	0.48%	0.43%	0.33%
	Gemini-Pro			3.42%	4.29%	8.15%	6.04%
	GPT-4V			8.12%	12.38%	**19.74%**	15.05%
Multimodal (SoM)	IDEFICS-80B-Instruct		Image + Caps + SoM	0.85%	0.95%	1.07%	0.99%
	CogVLM			0.00%	0.48%	0.43%	0.33%
	Gemini-Pro			3.42%	3.81%	7.73%	5.71%
	GPT-4V			**9.83%**	**17.14%**	19.31%	**16.37%**
Human Performance	-	-	Webpage	91.07%	87.10%	88.39%	88.70%

Table 3: Success rates of baseline LLM and VLM agents on `VisualWebArena`.

Zhou, Shuyan, et al. "Webarena: A realistic web environment for building autonomous agents." arXiv preprint arXiv:2307.13854 (2023).
Koh, Jing Yu, et al. "Visualwebarena: Evaluating multimodal agents on realistic visual web tasks." arXiv preprint arXiv:2401.13649 (2024).

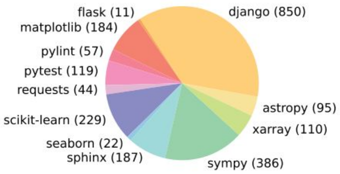
Agent Benchmarks with Device Control

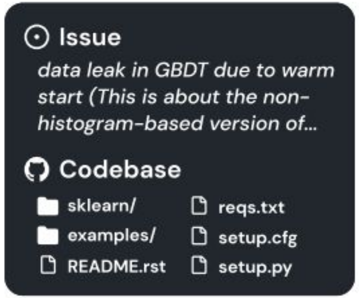
SWE-Bench

Goal: Evaluate VLM's ability to resolve github codebase issues

Environment: Clones of popular github repos and their resolved issues

Task Input: full repo code base + one issue





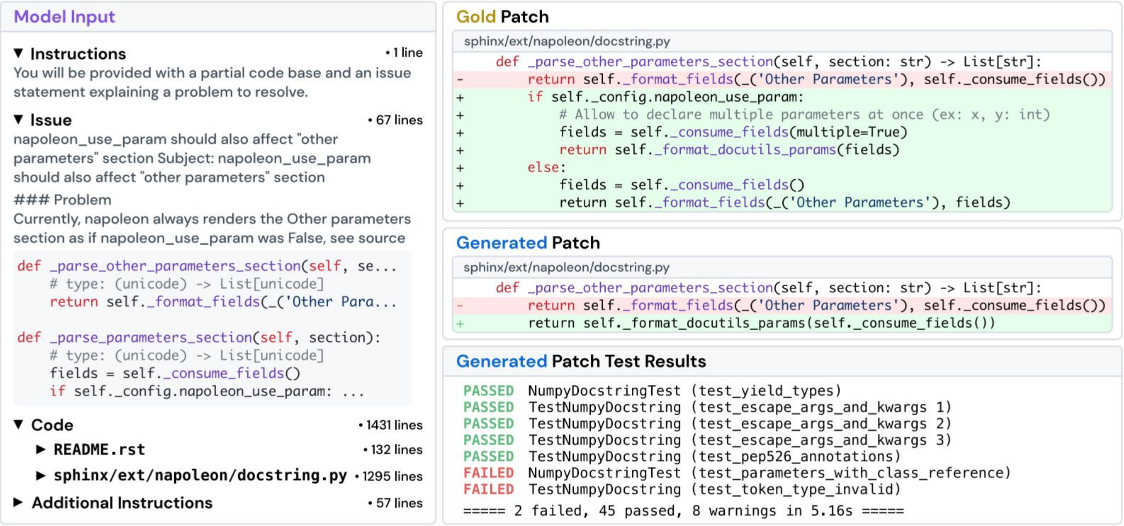
Evaluation: unit tests from pull request in that repo

Jimenez, Carlos E., et al. "Swe-bench: Can language models resolve real-world github issues?." arXiv preprint arXiv:2310.06770 (2023).

Agent Benchmarks with Device Control

SWE-Bench

Baselines: use retrieval models (e.g., BM25) to select relevant files, and prompt an LLM to generate the required code changes (i.e., code patch)



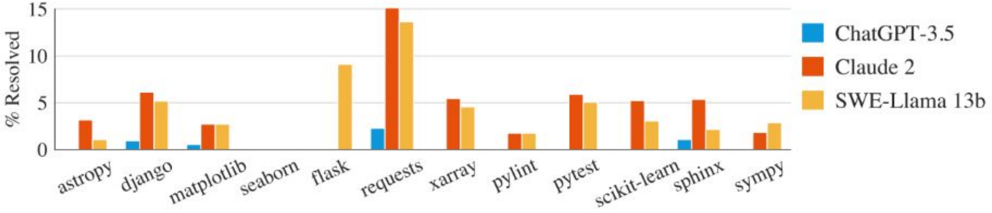
Jimenez, Carlos E., et al. "Swe-bench: Can language models resolve real-world github issues?." arXiv preprint arXiv:2310.06770 (2023).

Agent Benchmarks with Device Control

SWE-Bench

Results: current state-of-the-art VLMs significantly lag behind human

	SWE-bench		SWE-bench Lite	
Model	% Resolved	% Apply	% Resolved	% Apply
Claude 3 Opus	**3.79**	46.56	**4.33**	**51.67**
Claude 2	1.97	43.07	3.00	33.00
ChatGPT-3.5	0.17	26.33	0.33	10.00
GPT-4-turbo	1.31	26.90	2.67	29.67
SWE-Llama 7b	0.70	51.74	1.33	38.00
SWE-Llama 13b	0.70	**53.62**	1.00	38.00



Jimenez, Carlos E., et al. "Swe-bench: Can language models resolve real-world github issues?." arXiv preprint arXiv:2310.06770 (2023).

Agent Benchmarks with Device Control

OSWorld

Goal: Evaluate VLM's ability to control a computer (e.g., ubuntu)

Environment: Ubuntu hosted in a docker container/VMWare

Task Input: instruction + initial computer screenshot

Initial State	Task Instruction
	*Can you help me clean up my computer by getting rid of all the cookies that Amazon might have saved?*

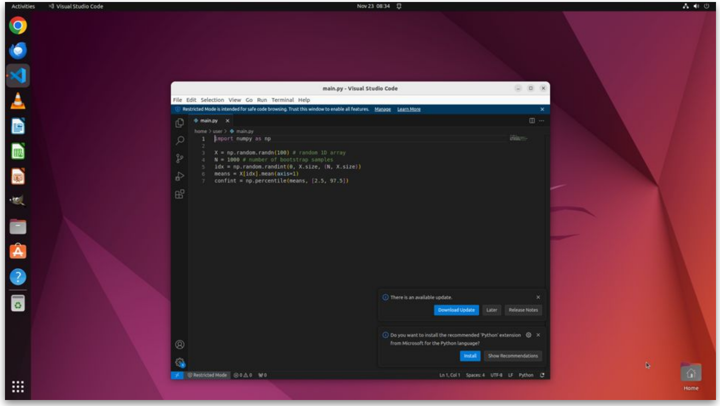
Evaluation: pre-defined scripts to check final state of the computer

Xie, Tianbao, et al. "Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments." arXiv preprint arXiv:2404.07972 (2024).

Agent Benchmarks with Device Control

OSWorld

Baselines: convert clickable buttons to their 2D coordinate on the screen, and prompt an VLM to output the coordinate to click next



screenshot

+

```
tag     name     text     class    description     position (top-left x&y) size (w&h)
label   Home     Home                    (1833, 1037)     (40, 17)
document-web     Welcome - Visual Studio Code    ""                   (448, 279)      (1024, 743)
section Explorer (Ctrl+Shift+E) ""                        (448, 279)      (48, 48)
section          î«°                      (460, 291)       (24, 24)
static  î«°      î«°                      (460, 291)       (24, 24)
section Search (Ctrl+Shift+F)    ""                       (448, 327)      (48, 48)
section          î©                     (460, 339)        (24, 24)
static  î©      î©                      (460, 339)        (24, 24)
section Source Control (Ctrl+Shift+G)    ""                       (448, 375)      (48, 48)
section          î©¨                      (460, 387)       (24, 24)
static  î©¨      î©¨                      (460, 387)       (24, 24)
section Run and Debug (Ctrl+Shift+D)     ""                       (448, 423)      (48, 48)
section          î©‘                      (460, 435)       (24, 24)
static  î©‘      î©‘                      (460, 435)       (24, 24)
section Extensions (Ctrl+Shift+X)        ""                       (448, 471)      (48, 48)
section          î«¦                      (460, 483)       (24, 24)
static  î«¦      î«¦                      (460, 483)       (24, 24)
push-button      Accounts        ""               (448, 904)      (48, 48)
section Accounts        ""              (448, 904)      (48, 48)
section          î©™                      (460, 916)       (24, 24)
static  î©™      î©™                      (460, 916)       (24, 24)
push-button      Manage  ""              (448, 952)      (48, 48)
section Manage  ""                       (448, 952)       (48, 48)
section          î‘                       (460, 964)       (24, 24)
static  î‘       î‘                       (460, 964)       (24, 24)
heading EXPLORER        EXPLORER                          (516, 279)      (54, 35)
static  EXPLORER        EXPLORER                          (516, 290)      (54, 12)
push-button      Views and More Actions...       ""                (717, 285)      (22, 23)
push-button      No Folder Opened Section        ""                (496, 314)      (255, 22)
section          îª´                      (498, 317)       (16, 16)
static  îª´      îª´                      (498, 317)       (16, 16)
heading NO FOLDER OPENED        NO FOLDER OPENED                  (516, 314)      (112, 22)
static  NO FOLDER OPENED        NO FOLDER OPENED                  (516, 319)      (112, 12)
```

(Simplified) Accessibility tree

Agent Benchmarks with Device Control

OSWorld

Results: current state-of-the-art VLMs significantly lag behind human

Inputs	Model	Success Rate (↑)					
		OS	Office	Daily	Profess.	**Workflow**	**Overall**
A11y tree	Mixtral-8x7B	12.50%	1.01%	4.79%	6.12%	0.09%	2.98%
	Llama-3-70B	4.17%	1.87%	2.71%	0.00%	0.93%	1.61%
	GPT-3.5	4.17%	4.43%	2.71%	0.00%	1.62%	2.69%
	GPT-4	20.83%	3.58%	25.64%	26.53%	2.97%	**12.24%**
	Gemini-Pro	4.17%	1.71%	3.99%	4.08%	0.63%	2.37%
	Gemini-Pro-1.5	12.50%	2.56%	7.83%	4.08%	3.60%	4.81%
	Qwen-Max	29.17%	3.58%	8.36%	10.20%	2.61%	6.87%
	GPT-4o	20.83%	6.99%	16.81%	16.33%	**7.56%**	11.36%
Screenshot	CogAgent	4.17%	0.85%	2.71%	0.00%	0.00%	1.11%
	GPT-4V	12.50%	1.86%	7.58%	4.08%	**6.04%**	5.26%
	Gemini-ProV	8.33%	3.58%	6.55%	16.33%	2.08%	**5.80%**
	Gemini-Pro-1.5	12.50%	6.99%	2.71%	6.12%	3.60%	5.40%
	Claude-3-Opus	4.17%	1.87%	2.71%	2.04%	2.61%	2.42%
	GPT-4o	8.33%	3.58%	6.07%	4.08%	5.58%	5.03%
Screenshot + A11y tree	CogAgent	4.17%	0.85%	2.71%	0.62%	0.09%	1.32%
	GPT-4V	16.66%	6.99%	24.50%	18.37%	4.64%	**12.17%**
	Gemini-ProV	4.17%	4.43%	6.55%	0.00%	1.52%	3.48%
	Gemini-Pro-1.5	12.50%	3.58%	7.83%	8.16%	1.52%	5.10%
	Claude-3-Opus	12.50%	3.57%	5.27%	8.16%	1.00%	4.41%
	GPT-4o	41.67%	6.16%	12.33%	14.29%	**7.46%**	11.21%
Set-of-Mark	CogAgent	4.17%	0.00%	2.71%	0.00%	0.53%	0.99%
	GPT-4V	8.33%	8.55%	22.84%	14.28%	**6.57%**	**11.77%**
	Gemini-ProV	4.17%	1.01%	1.42%	0.00%	0.63%	1.06%
	Gemini-Pro-1.5	16.67%	5.13%	12.96%	10.20%	3.60%	7.79%
	Claude-3-Opus	12.50%	2.72%	14.24%	6.12%	4.49%	6.72%
	GPT-4o	20.83%	3.58%	3.99%	2.04%	3.60%	4.59%
Human Performance		75.00%	71.79%	70.51%	73.47%	73.27%	72.36%

Agent Benchmarks with Device Control

Android in the Wild

Goal: Evaluate VLM's ability to resolve control apps in mobile phones

Environment: Android device emulator

Task Input: instruction + initial phone screenshot

Turn on bluetooth scan

Evaluation: checks if VLM's generated actions match ground-truth actions

Rawles, Christopher, et al. "Androidinthewild: A large-scale dataset for android device control." *Advances in Neural Information Processing Systems* 36 (2024).

Agent Benchmarks with Device Control

Android in the Wild

Baselines: combine multiple models trained to output actions such as click, swipe, etc; and also prompt an VLM to solve the task

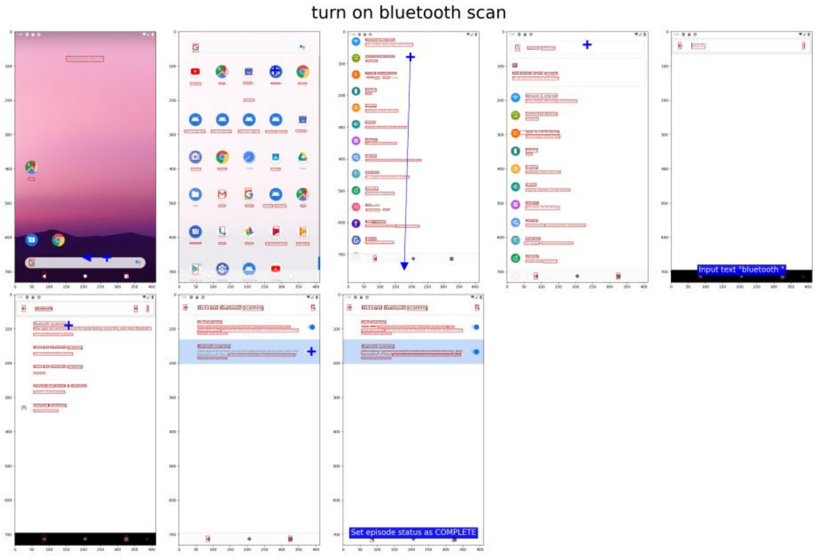


Figure 4: Example episode from the dataset.

Agent Benchmarks with Device Control

Android in the Wild

Results: can reach 70%+ performance when trained,
but existing LLMs can only reach 40% without training

Model	Standard	Out-of-domain generalization			
		Version	Subject	Verb	Domain
BC-single	68.7	59.2	64.2	66.4	52.2
BC-history	**73.1**	**63.2**	**68.5**	**70.4**	**59.7**
LLM-0 [43]	30.9 [25.6, 36.6]	31.6 [26.3, 37.3]	33.7 [28.2, 39.5]	32.6 [27.3, 38.4]	25.3 [20.4, 30.8]
LLM-hist-5-CoT	39.6 [33.9, 45.5]	29.5 [24.3, 35.1]	44.4 [38.6, 50.4]	41.7 [35.9, 47.6]	35.8 [30.2, 41.6]

Rawles, Christopher, et al. "Androidinthewild: A large-scale dataset for android device control." *Advances in Neural Information Processing Systems* 36 (2024).

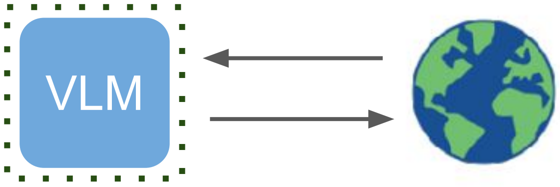
Summary

- ❖ Visual Language Agents for device control
    - Tremendous practical values
    - Scalable environment
- ❖ Challenges
    - Difficult to scale automatic evaluation
    - VLMs or RL agents cannot (yet) solve it

2

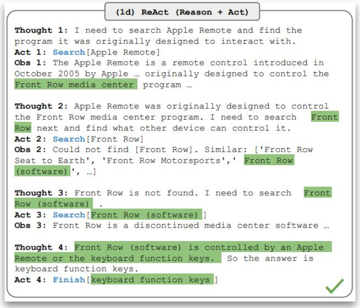
Building (visual) language agents for device control

Inference/training methods to improve agent performance

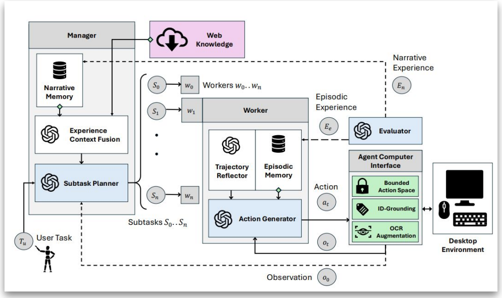


Agentic Methods to Improve Task Performance

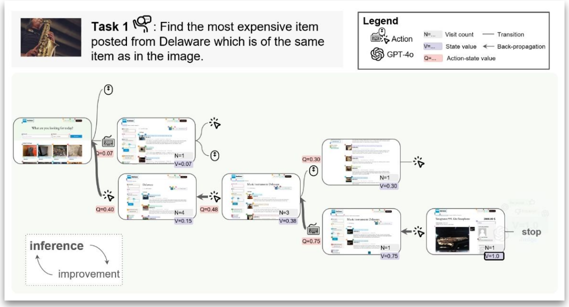
Recent inference-time methods (non-exhaustive)



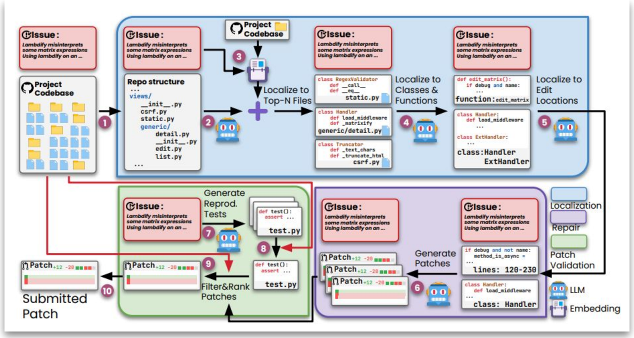
Prompting VLMs to Reason



Prompting VLMs to Decompose tasks



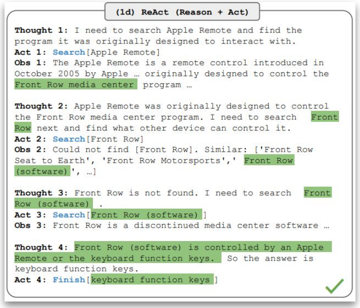
Augmenting with Search Algorithms



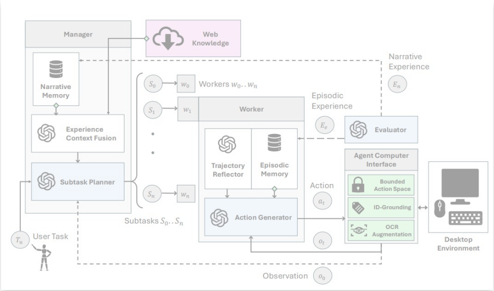
Augmenting with Tools

Agentic Methods to Improve Task Performance

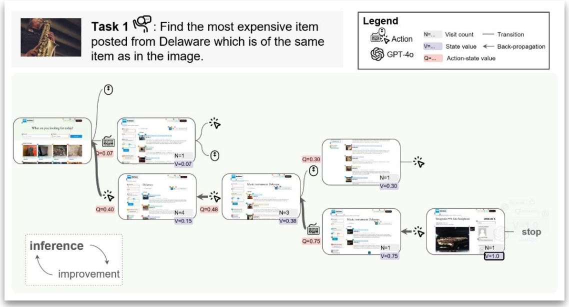
Recent inference-time methods (non-exhaustive)



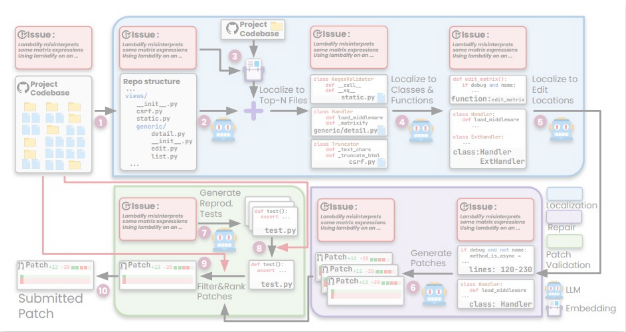
Prompting VLMs to Reason



Prompting VLMs to Decompose tasks



Augmenting with Search Algorithms

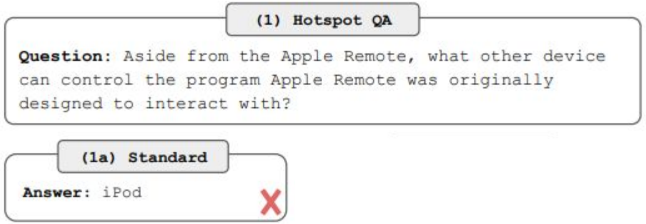


Augmenting with Tools

Agentic Methods to Improve Task Performance

Prompting VLMs to Reason

Main Idea: "force" a (V)LM to reason about the current environment **before** generating an action

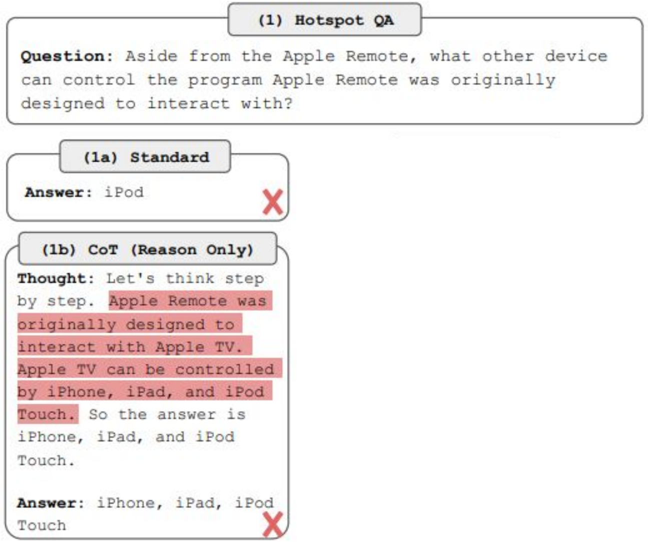


Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." *arXiv preprint arXiv:2210.03629* (2022).

Agentic Methods to Improve Task Performance

Prompting VLMs to Reason

Main Idea: “force” a (V)LM to reason about the current environment **before** generating an action

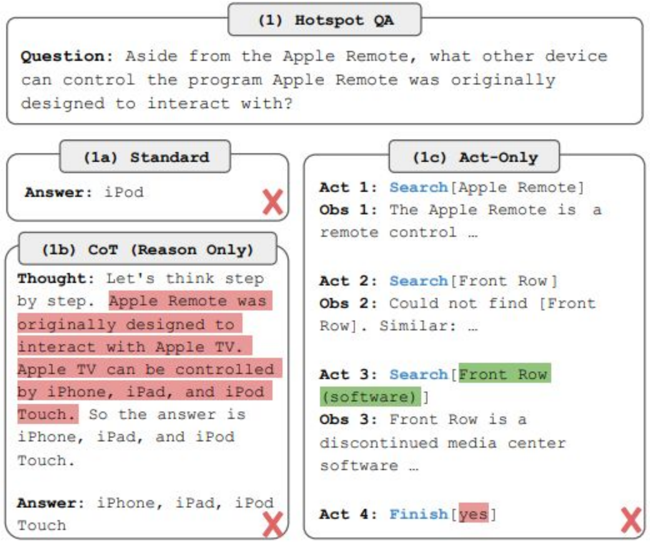


Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." *arXiv preprint arXiv:2210.03629* (2022).

Agentic Methods to Improve Task Performance

Prompting VLMs to Reason

Main Idea: “force” a (V)LM to reason about the current environment **before** generating an action

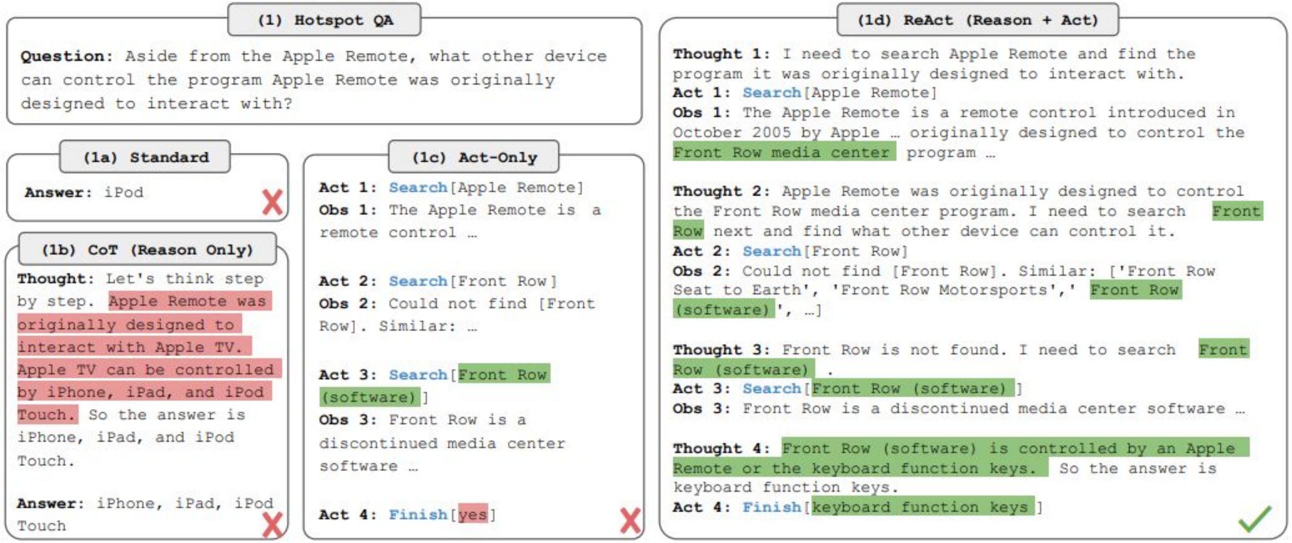


Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." *arXiv preprint arXiv:2210.03629* (2022).

Agentic Methods to Improve Task Performance

Prompting VLMs to Reason

Main Idea: "force" a (V)LM to reason about the current environment **before** generating an action



Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." *arXiv preprint arXiv:2210.03629* (2022).

Agentic Methods to Improve Task Performance

Prompting VLMs to Reason

Main Results:
- combined with methods such as self-consistency (SC), ReACT can significantly **improve performance**
- performance improves when **more compute (SC trials) is allocated**

Prompt Method[a]	**HotpotQA** (EM)	**Fever** (Acc)
Standard	28.7	57.1
CoT (Wei et al., 2022)	29.4	56.3
CoT−SC (Wang et al., 2022a)	33.4	60.4
Act	25.7	58.9
ReAct	27.4	60.9
CoT−SC → ReAct	34.2	**64.6**
ReAct→ CoT−SC	**35.1**	62.0
Supervised SoTA[b]	67.5	89.5

Table 1: PaLM-540B prompting results on HotpotQA and Fever.

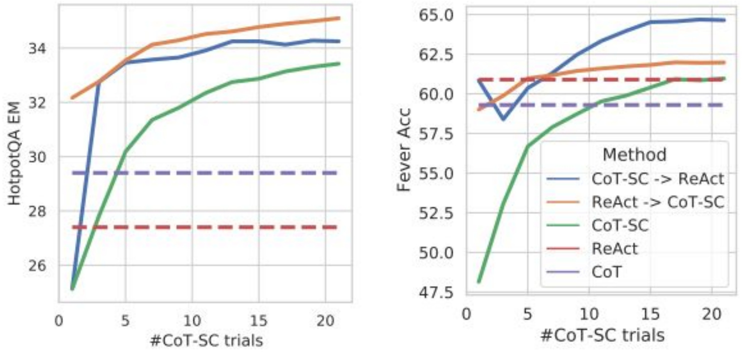
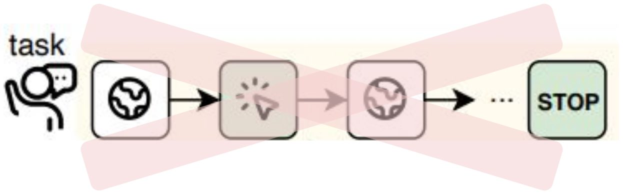


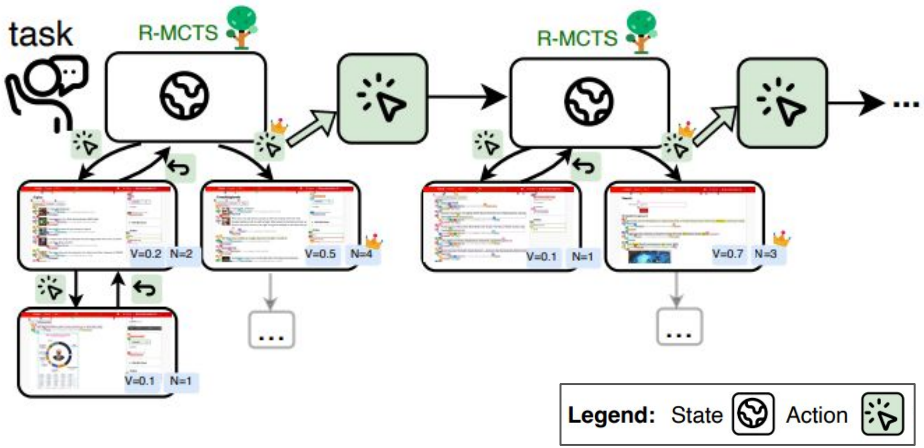
Figure 2: PaLM-540B prompting results with respect to number of CoT−SC samples used.

Agentic Methods to Improve Task Performance

Augmenting Agent with Search Algorithms

Main Idea: "force" a (V)LM to make informed decisions **after** interacting/exploring the environment



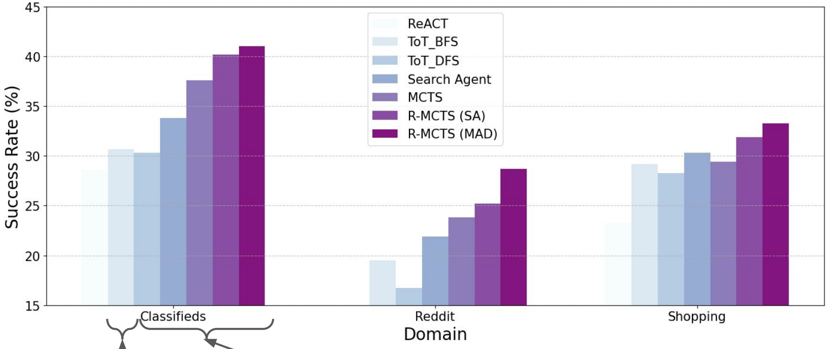


Yu, Xiao, et al. "ExACT: Teaching AI Agents to Explore with Reflective-MCTS and Exploratory Learning." *arXiv preprint arXiv:2410.02052* (2024).

Agentic Methods to Improve Task Performance

Augmenting Agent with Search Algorithms

Main Results:
- Search methods such as R-MCTS can significantly **improve performance**
- Scales when **more search budget/tokens is allocated**



non-search

augmented with search

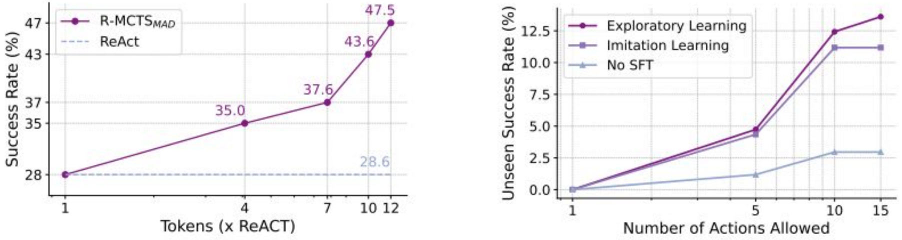


Figure 1: Our work yields compute scaling of GPT-4o with R-MCTS (left) and fine-tuned GPT-4o (right), for both training and testing, respectively. Left is evaluated on all 234 tasks from Classifieds in VisualWebArena, and right is evaluated on 169 unseen tasks from Classifieds.

Yu, Xiao, et al. "ExACT: Teaching AI Agents to Explore with Reflective-MCTS and Exploratory Learning." *arXiv preprint arXiv:2410.02052* (2024).

Agentic Methods to Improve Task Performance

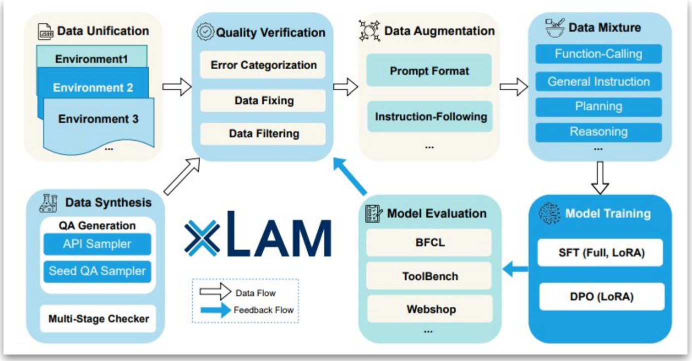
Augmenting Agent with Search Algorithms

Demos: you can find more info/examples at https://agent-e3.github.io/ExACT/

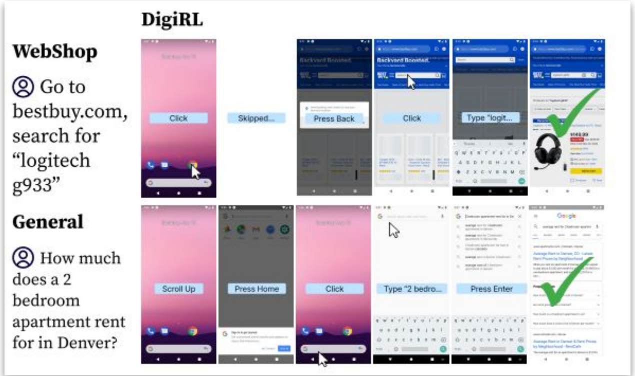
Yu, Xiao, et al. "ExACT: Teaching AI Agents to Explore with Reflective-MCTS and Exploratory Learning." *arXiv preprint arXiv:2410.02052* (2024).

Agentic Methods to Improve Task Performance

Recent training methods (non-exhaustive)



Supervised Learning with Human/Synthetic Data

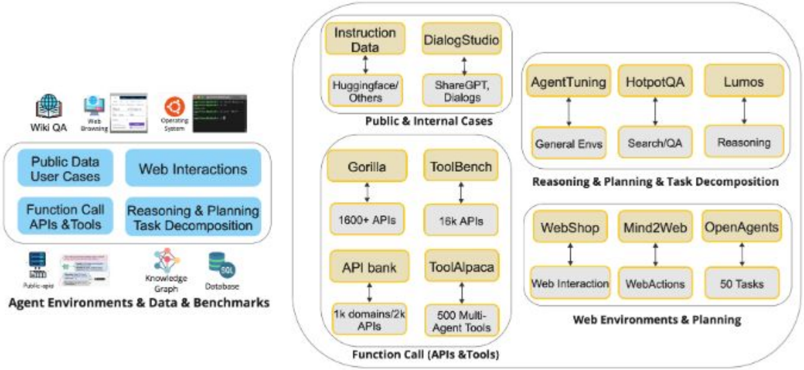


Reinforcement Learning with Simulated Evaluation

Zhang, Jianguo, et al. "xlam: A family of large action models to empower ai agent systems." *arXiv preprint arXiv:2409.03215* (2024).
Bai, Hao, et al. "Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning." *arXiv preprint arXiv:2406.11896* (2024).

Agentic Methods to Improve Task Performance

Recent training methods (non-exhaustive)



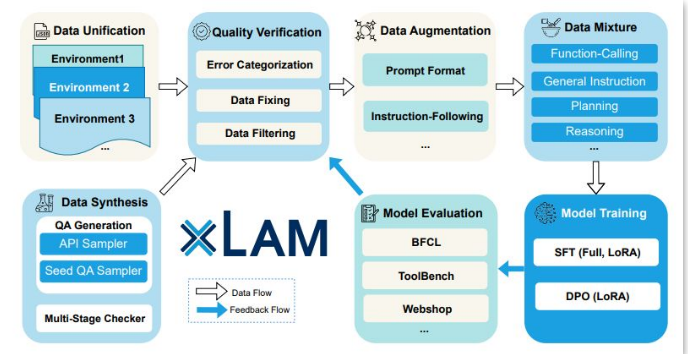
Supervised Learning with Human/Synthetic Data

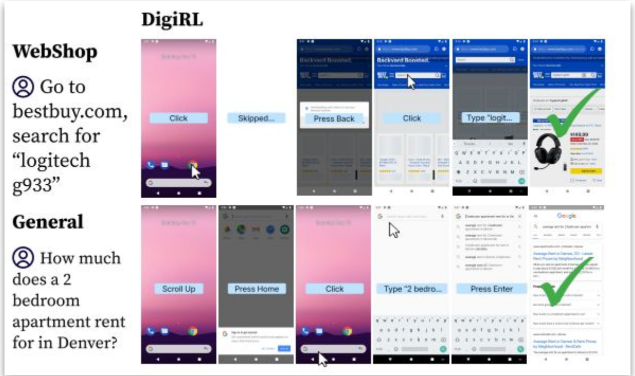
Reinforcement Learning with Simulated Evaluation

Zhang, Jianguo, et al. "xlam: A family of large action models to empower ai agent systems." *arXiv preprint arXiv:2409.03215* (2024).
Bai, Hao, et al. "Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning." *arXiv preprint arXiv:2406.11896* (2024).

Agentic Methods to Improve Task Performance

Recent training methods (non-exhaustive)





Challenge 1: hard to gather labeled human demonstrations in large scale
Challenge 2: hard to accurately evaluate/estimate whether a task is solved

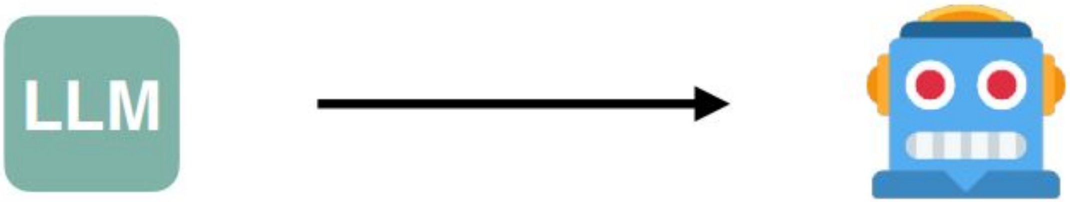
Summary

- Inference/training methods to improve agent performance
  - Search-based inference methods are scalable
  - Many methods to improve performance

- Challenges
  - Difficult to scale data collection (for training)
  - Difficult to scale automatic evaluation (for training + inference)

Final Remarks from Shunyu Yao's PhD Thesis Defense

The most powerful neural networks ever built shouldn't just answer questions or draft emails.



They should be used to automate every aspect of our life, society, and science.

Attendance Question

Which of the following benchmarks have we **NOT** talked about in this lecture:

- OSWorld
- WebArena
- AppWorld
- Android in the Wild