Part III: Applications, Data, and Evaluation

Tao Yu

Agenda

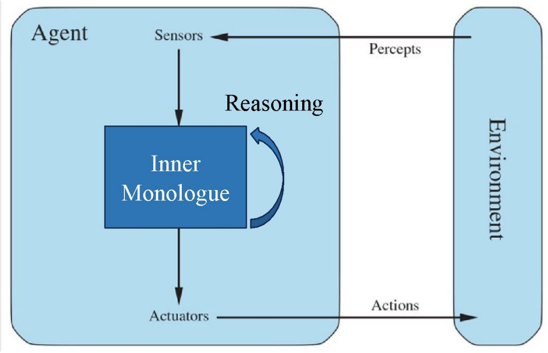
- Agent applications
    - in digital world
    - in physical world
- Agent data
    - via human demonstrations
    - via synthesis and simulation
    - via internet-scale data
- Agent evaluation
    - via benchmarks
    - via LLMs/VLMs
    - via crowdsourcing

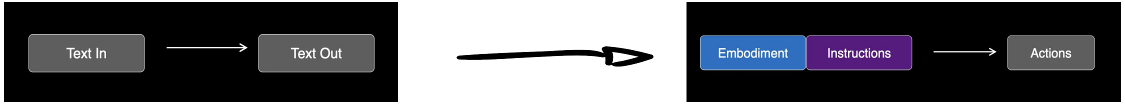
Agent applications

Agent applications

Agent applications by embodiment

- Digital world
  - Coding agents
  - Gaming agents
  - Mobile agents
  - Web/app agents
  - Computer agents
- Physical world
  - Robotics





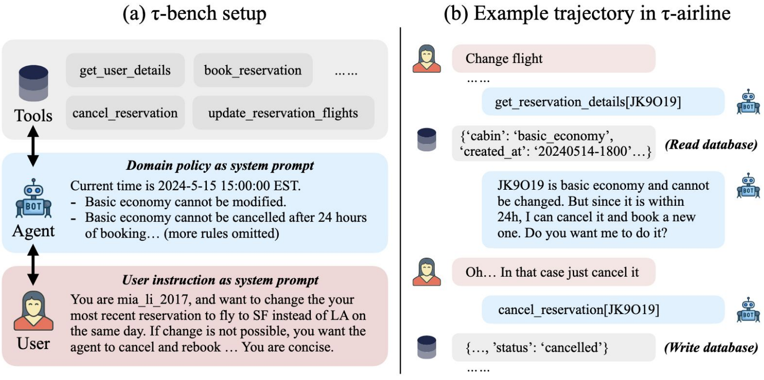
Agent applications

in digital world

Coding agents

API/functional calling for tool use

- Environment: software systems such as databases, app/web services…
- Observation space: API docs, system info, error messages and logs…
- Action space: function calls, error handling routines…

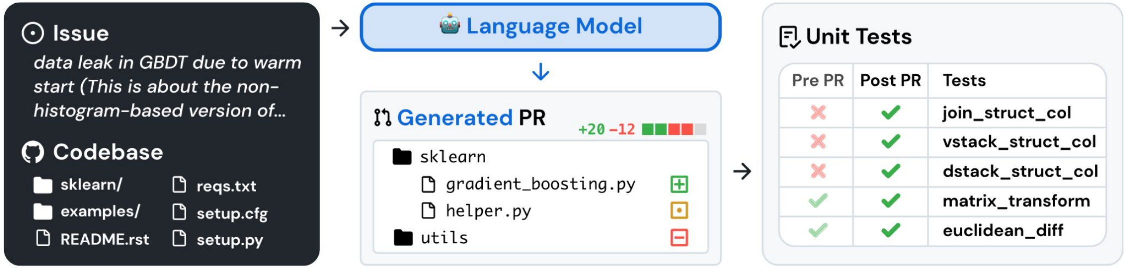


(a) An orders database entry in τ-retail.
(b) An API tool in τ-retail.
(c) Domain policy excerpts in τ-retail.
(d) User instruction ensures only one possible outcome.

Coding agents

Project-level coding tasks

- Environment: project code repos, filesystems, IDEs…
- Observation space: code files, exe outputs, docs, errors, commit history…
- Action space: code edits, file search/view, test updates…



SWE-bench: Can Language Models Resolve Real-World GitHub Issues?, (Jimenez et al., 2023)
Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows, (Lei et al., 2024)
MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering, (Chan et al., 2024)

Agents for software development

Coding agents won't be our focus in this tutorial.

Agents for Software Development, (Neibig et al., 2024)

Gaming agents

Digital games

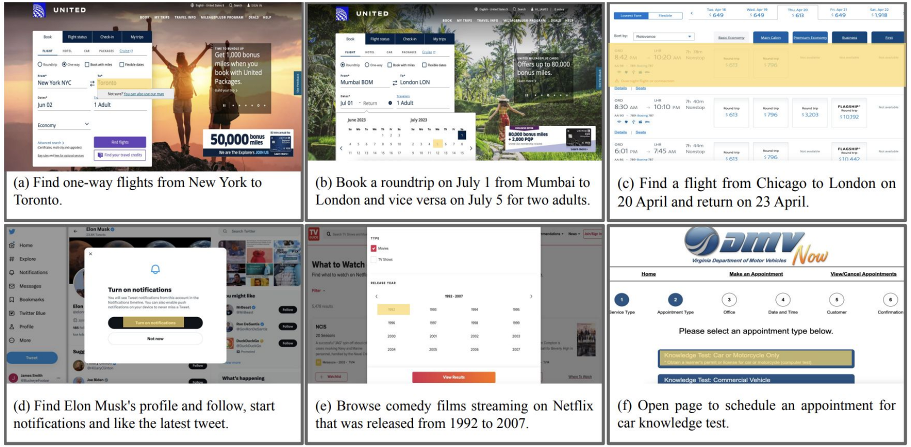
- Environment: game worlds/levels...
- Observation space: screenshots of game states, inventory, location...
- Action space: game controls (e.g., drop, move, attack, resource management...)

MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. (Fan et al., 2022)

Web/app agents

Web/app use

- Environment: web browsers/apps
- Observation space: screenshots, DOM trees, HTML, historical actions…
- Action space: browser/app controls (e.g., click, type, scroll, drag, hover…)



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

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

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

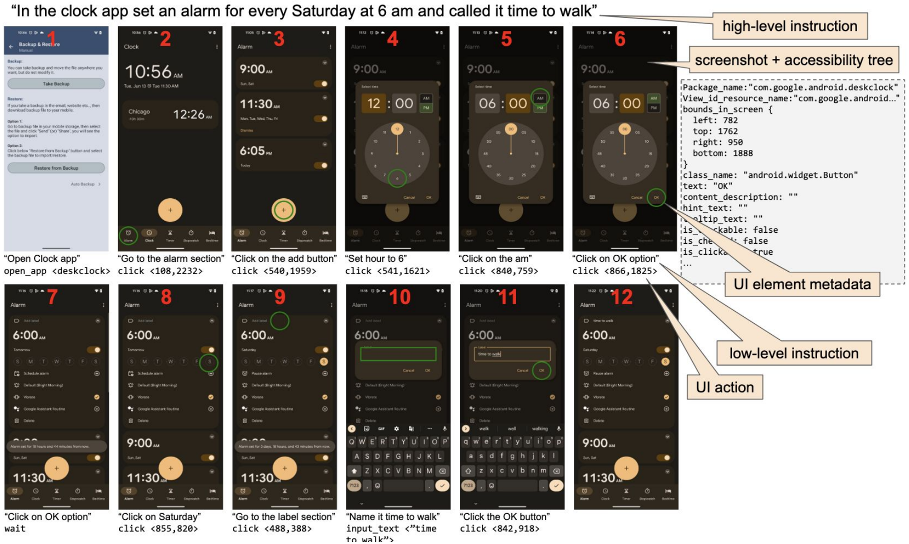
(f) Open page to schedule an appointment for car knowledge test.

World of Bits: An Open-Domain Platform for Web-Based Agents, (Shi et al., 2017)
Mind2Web: Towards a Generalist Agent for the Web, (Deng et al., 2023)
WebArena: A Realistic Web Environment for Building Autonomous Agents, (Zhou et al., 2023)

Mobile agents

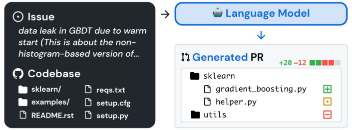
Mobile use

- Environment: mobile device systems
- Observation space: screenshots, a11y trees, HTML, historical actions…
- Action space: mobile controls (e.g., tap, type, swipe…)



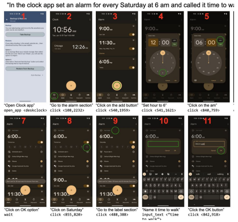
On the Effects of Data Scale on Computer Control Agents, (Li et al., 2024)

Universal digital environment

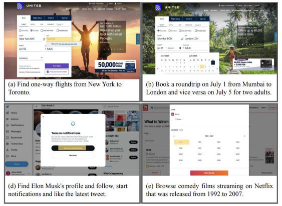


Coding

Gaming

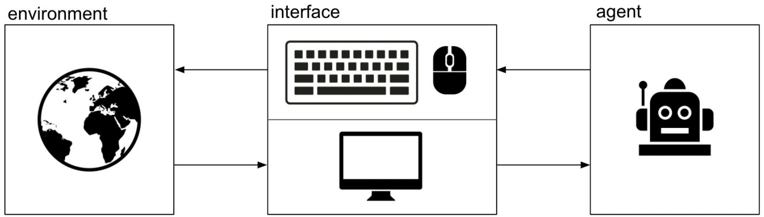


Mobile



Web/apps

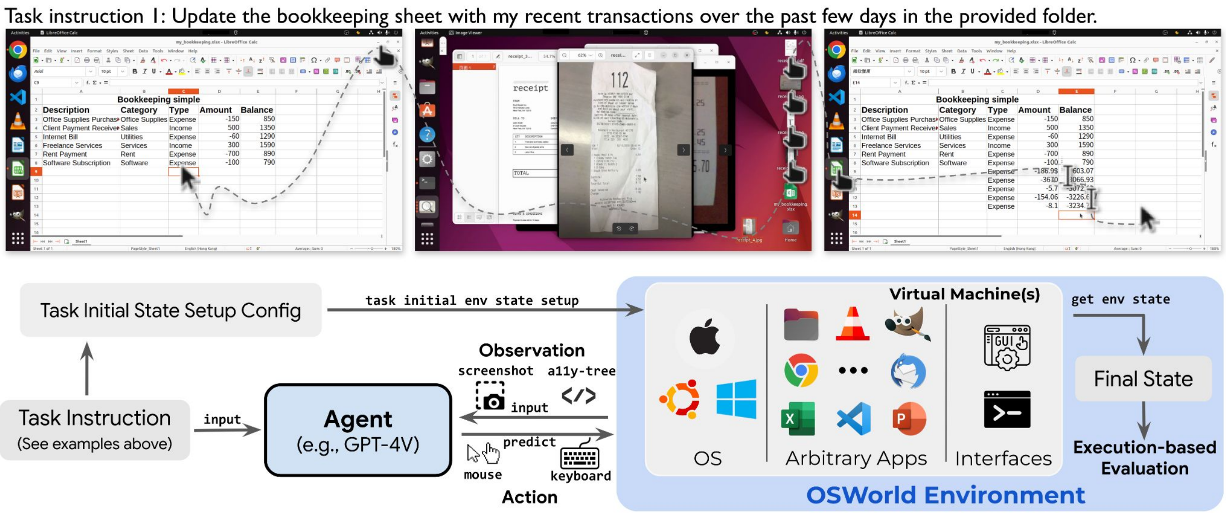
Can we study all digital AI agents in a **single** environment with **unified** observation and action spaces?



Computer use agents

Computer use for universal digital tasks

- Environment: desktop operating systems
- Observation space: desktop screenshots, a11y trees, historical actions…
- Action space: keyboard/mouse controls (e.g., click, type, drag, shortcuts)



OpenAI Universe, (OpenAI, 2016)
OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments, (Xie et al., 2024)

Computer use agents

Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku

Oct 22, 2024 • 5 min read

Category	Claude 3.5 Sonnet (New) - 15 steps		Claude 3.5 Sonnet (New) - 50 steps		Human Success Rate [3]
	Success Rate	95% CI	Success Rate	95% CI	
OS	54.2%	[34.3, 74.1]%	41.7%	[22.0, 61.4]%	**75.00%**
Office	7.7%	[2.9, 12.5]%	17.9%	[11.0, 24.8]%	**71.79%**
Daily	16.7%	[8.4, 25.0]%	24.4%	[14.9, 33.9]%	**70.51%**
Professional	24.5%	[12.5, 36.5]%	40.8%	[27.0, 54.6]%	**73.47%**
Workflow	7.9%	[2.6, 13.2]%	10.9%	[4.9, 17.0]%	**73.27%**
Overall	14.9%	[11.3, 18.5]%	22%	[17.8, 26.2]%	**72.36%**

Anthropic recent computer use agent results of OSWorld

Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, (Anthropic, 2024)
Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet, (Anthropic, 2024)

Agent applications
in physical world

Robotic agents

Robotics for physical interaction

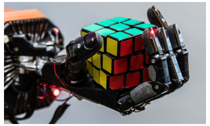
- Environment: physical world spaces
- Observation space: visual input, sensor readings, physical states, proprioception…
- Action space: motor controls (e.g., move, grasp, manipulate…)

Project GR00T: A Blueprint for Generalist Robotics, (Fan et al., 2024)

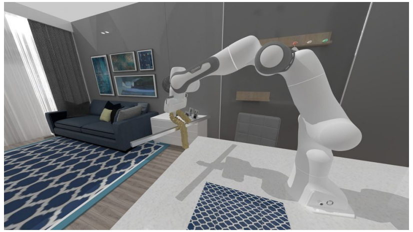
Agent application overview

Physical agent tasks

- Observation: sensor data streams
- Control complexity: high



- Task distribution: more concentrated, natural

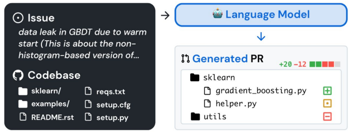


- Data collection: very hard (simulation)
- Evaluation: very hard (simulation)
- Deployment: complex (sim2real gap)

Digital agent tasks

- Observation: screen/UI states
- Control complexity: Low

- Task distribution: long tail, reasoning-intensive



Coding

Professional workflows

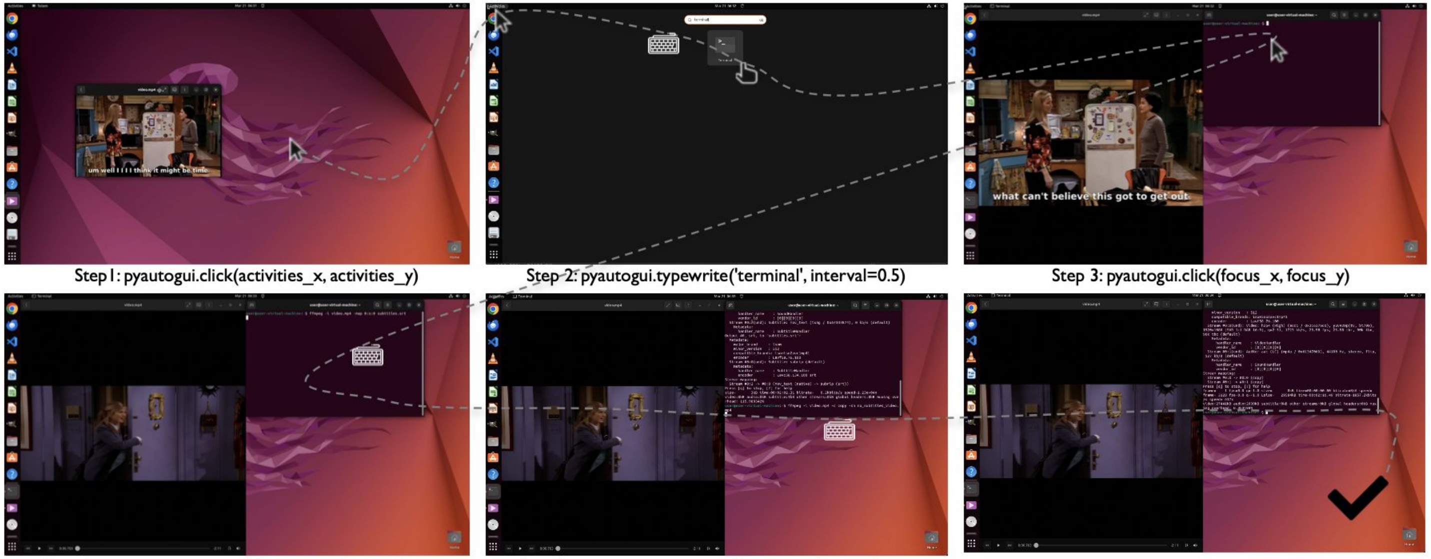
- Data collection: hard (real env)
- Evaluation: hard (real env)
- Deployment: easy (no sim2real gap)

Agent data

Agent data example

Task goal aligned trajectories (observation-action pairs)

Task Instruction: I downloaded an episode of Friends to practice listening, but I don't know how to remove the subtitles. Please help me remove the subtitles from the video and export it as "subtitles.srt" and store it in the same directory as the video.



Step 1: pyautogui.click(activities_x, activities_y)

Step 2: pyautogui.typewrite('terminal', interval=0.5)

Step 3: pyautogui.click(focus_x, focus_y)

Step 4: pyautogui.typewrite('ffmpeg -i video.mp4 -map 0:s:0 subtitles.srt', interval=0.5)

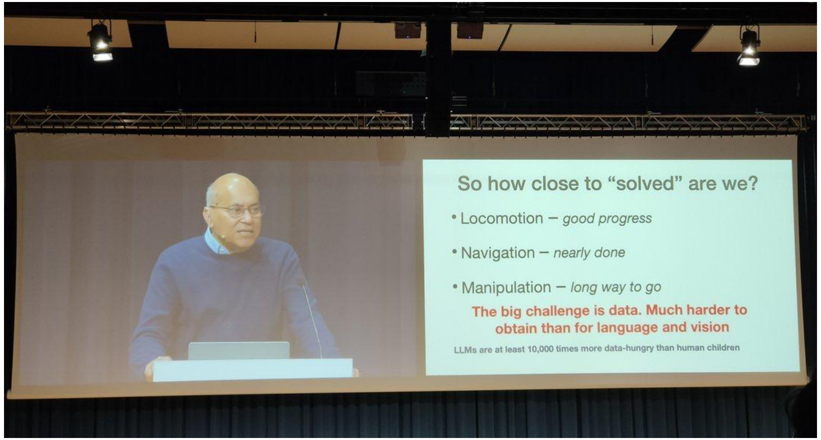
Step 5: pyautogui.typewrite('ffmpeg -i video.mp4 -c copy -sn no_subtitles_video.mp4', interval=0.5)

Step 6: Done

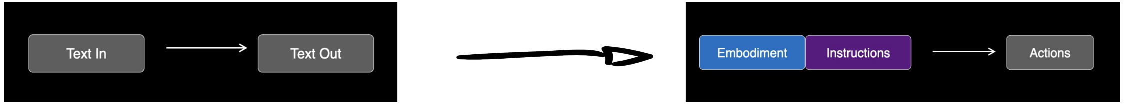
Agent data: a big challenge!

Agent data is hard to get directly from internet-scale text and videos due to **embodiment**.

- Complex data collection infrastructure
- Complex observation-action interaction in diverse environments
- Goal aligned trajectories



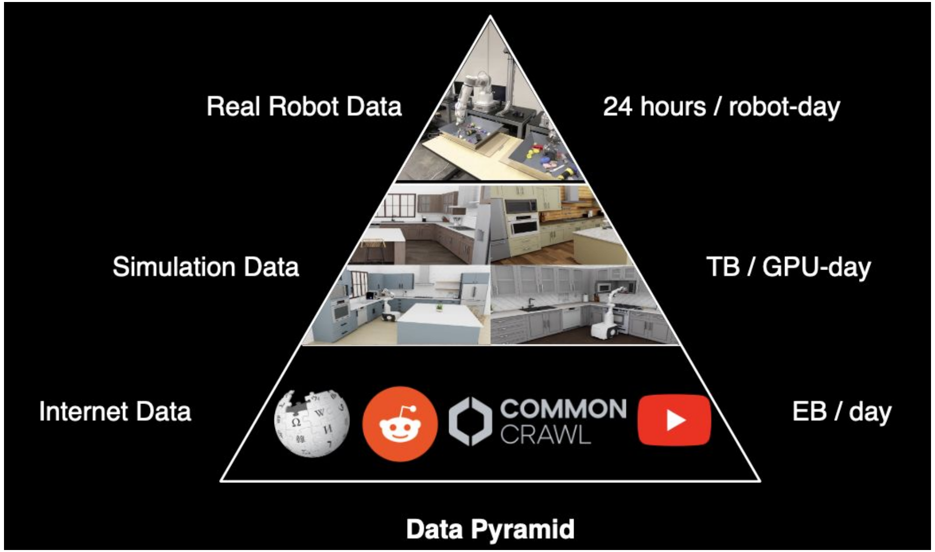
Jitendra's talk @ CoRL 2024



Scaling agent data

Scaling agent data through

- Human demonstrations
- Synthesis/simulation
- Internet-scale data



Project GR00T: A Blueprint for Generalist Robotics, (Fan et al., 2024)

Agent data
via human demonstrations

Agent data challenge 1: hard to collect

Human demonstration pipeline

- Task definition
- Infrastructure setup
- Task initial environment config
- Human demonstration recording
- Data verification

Agent data: not yet at scale

Data source	Platform	Inner Monologue	Avg. Steps	#Trajectory
MM-Mind2Web (Zheng et al., 2024a)	Website	Generated	7.7	1,009
GUIAct (Chen et al., 2024a)	Website	Generated	6.7	2,482
MiniWoB++ (Zheng et al., 2024b)	Website	Generated	3.6	2,762
AitZ (Zhang et al., 2024b)	Mobile	Original	6.0	1,987
AndroidControl (Li et al., 2024d)	Mobile	Original	5.5	13,594
GUI Odyssey (Lu et al., 2024)	Mobile	Generated	15.3	7,735
AMEX (Chai et al., 2024)	Mobile	Generated	11.9	2,991
AitW (Rawles et al., 2024b)	Mobile	Generated	8.1	2,346
Total				35K

Existing digital agent data

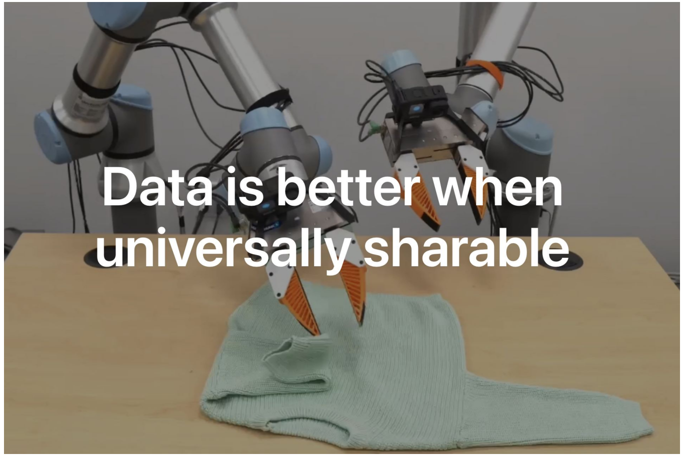
Feature	RoboCasa	AI2-THOR	Habitat 2.0	iGibson 2.0	RLBench	Behavior-1K	robomimic	ManiSkill 2	OPTIMUS	LIBERO	MimicGen
Mobile Manipulation	✓	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓
Room-Scale Scenes	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
Realistic Object Physics	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
AI-generated Tasks	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
AI-generated Assets	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Photorealism	✓	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗
Cross-Embodiment	✓	✓	✗	✓	✗	✓	✗	✗	✓	✗	✓
Num Tasks	100	-	3	6	100	1000	8	20	10	130	12
Num Scenes	120	-	1	15	1	50	3	-	4	20	1
Num Object Categories	153	-	46	-	28	1265	-	-	-	x	-
Num Objects	2509	3578	169	1217	28	5215	15	2144	72	x	40
Human Data	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
Machine-Generated Data	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓
Num Trajectories	100K+	-	-	-	-	0	6K	30K	245K	5K	50K

Existing robotic agent data

Agent data challenge 2: hard to share

Heterogeneous agent data formats

- Data from various platforms and embodied environments produces different observation and action spaces
- Make data merging and standardization difficult, hindering development of general-purpose agents

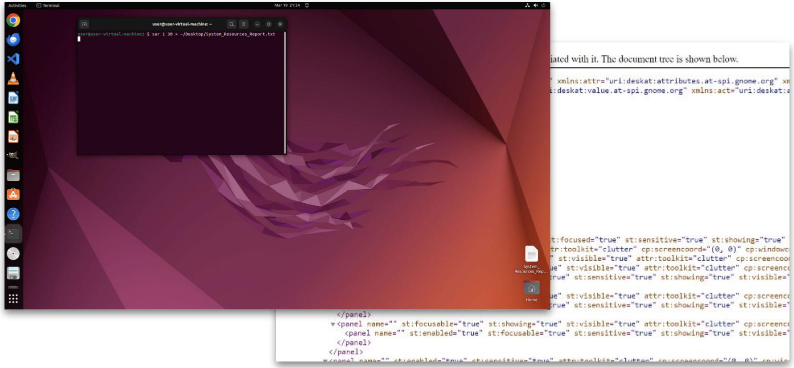


https://umi-data.github.io

Unifying digital agent data

Digital agent data unification

- Observation: screenshots (or a11y trees, but not reliable)
- Actions: universal computer mouse and keyboard controls



Observation: screenshots/a11y trees

Category	Action Space

Basic Actions	`pyautogui.moveTo(x, y)`
	`pyautogui.click(x, y)`
	`pyautogui.write('text')`
	`pyautogui.press('enter')`
	`pyautogui.hotkey('ctrl', 'c')`
	`pyautogui.scroll(200)`
	`pyautogui.dragTo(x, y)`

Pluggable Actions	`browser.select_option(x, y, value)`
	`mobile.swipe(from, to)`
	`mobile.home()`
	`mobile.back()`
	`mobile.open_app(name)`
	`terminate(status)`
	`answer(text)`

...	...

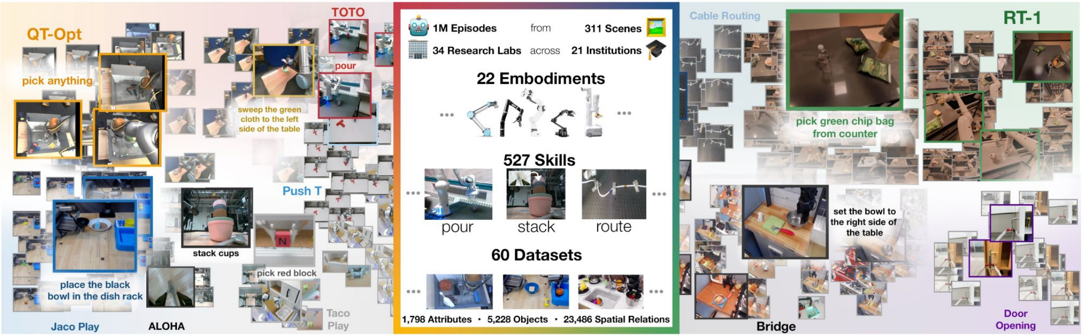
Actions: `pyautogui` computer control actions with pluggable actions

Unifying robotic agent data

Unifying data from diverse robotic platforms and sensors to enable large-scale agent training.

Open X-Embodiment: Robotic Learning Datasets and RT-X Models

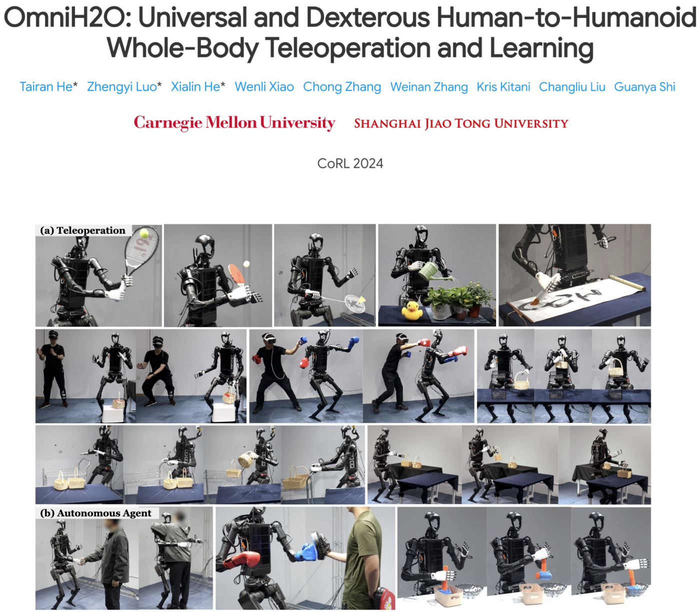
Open X-Embodiment Collaboration
(hover to display full author list)



Open X-Embodiment: Robotic Learning Datasets and RT-X Models, (Open X-Embodiment, 2023)
1st Workshop on X-Embodiment Robot Learning @ CoRL 2024

Unifying robotic agent data collection infrastructure

Simplifying and unifying robotic data collection hardwares

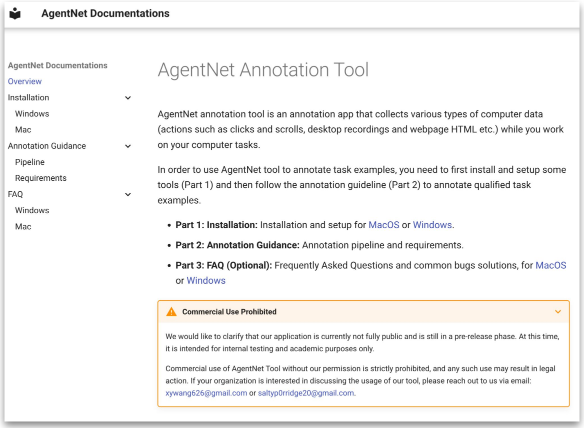


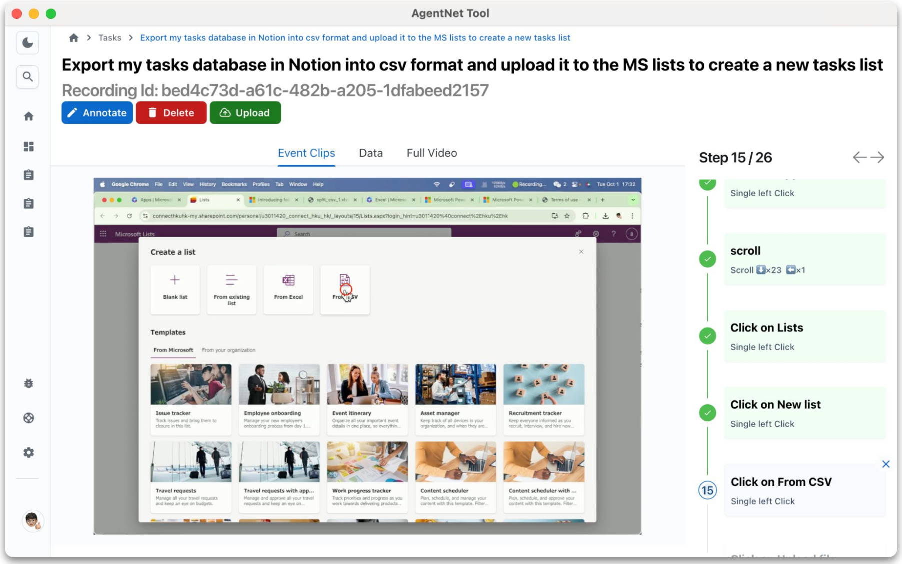
Universal Manipulation Interface In-The-Wild Robot Teaching Without In-The-Wild Robots, (Chi et al., 2024)
OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning, (He et al., 2024)
OpenVLA: An Open-Source Vision-Language-Action Model, (Kim et al., 2024)
Octo: An Open-Source Generalist Robot Policy, (Octo Team, 2024)

Unifying digital agent data collection infrastructure

Unifying digital agent data collection with our AgentNet tool

- Universal platform for digital agent data collection and verification in a unified data format





https://junliwang.tech/agentnet-docs

Human demonstration is hard to scale

Challenges in scaling human demonstration data collection

- Expensive and complex infrastructure setup
- Expert time & cost
- Task coverage

Possible solutions

- Data synthesis or simulation
- Leveraging internet-scale data

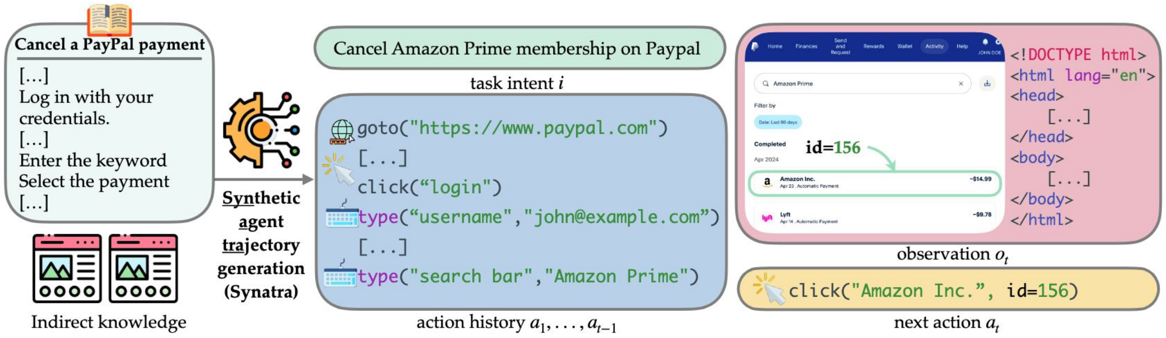
Agent data
via synthesis and simulation

Synthesizing digital agent data

Converting online tutorials into direct training demonstrations, making human-oriented instruction materials usable for training AI systems

Synatra: Turning Indirect Knowledge into Direct Demonstrations for Digital Agents at Scale

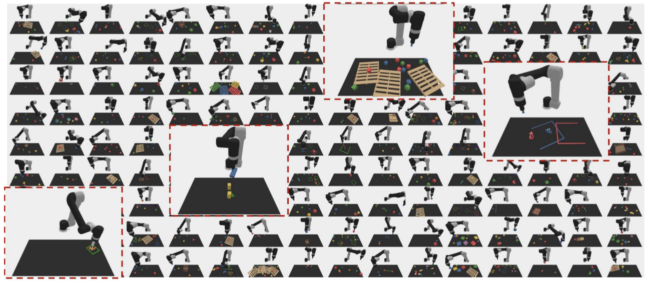
Tianyue Ou[1], Frank F. Xu[1], Aman Madaan [1],
Jiarui Liu[1], Robert Lo[1], Abishek Sridhar[1],
Sudipta Sengupta[2] Dan Roth[2] Graham Neubig[1] Shuyan Zhou[1]
[1]Carnegie Mellon University, [2]Amazon AWS AI

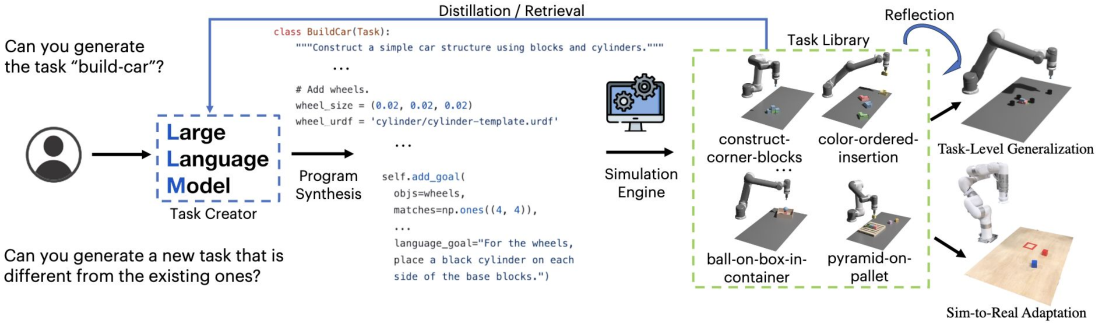


Synatra: Turning Indirect Knowledge into Direct Demonstrations for Digital Agents at Scale. (Ou et al., 2024)
BAGEL: Bootstrapping Agents by Guiding Exploration with Language. (Murty et al., 2024)
Agent Workflow Memory. (Wang et al., 2024)

Scaling robotic data via simulation

Generating simulation environments and expert demonstrations by leveraging LLM’s grounding and coding ability.





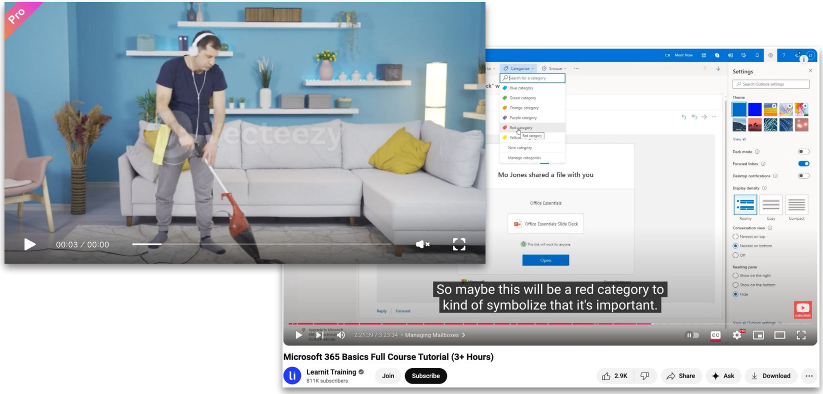
GenSim: Generating Robotic Simulation Tasks via Large Language Models. (Wang et al., 2023)
RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. (Nasiriany et al., 2024)
SOAR: Autonomous Improvement of Instruction Following Skills via Foundation Models. (Zhou et al., 2024)

Synthesizing good agent data is still challenging

Challenges in agent data synthesis and simulation

- Limited foundation model capabilities
- World knowledge or exploration limitations
- Sim2real gap

Possible solution: leveraging internet-scale human demonstration video data?

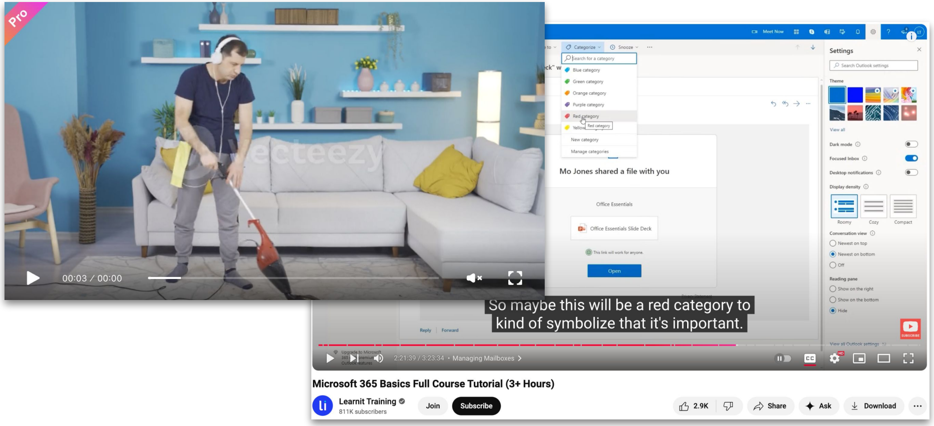


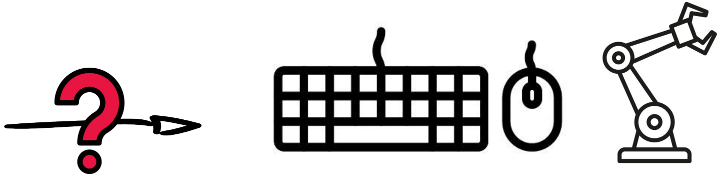
Agent data

via internet-scale data

Internet-scale data for agent learning

Numerous videos exist online showing humans demonstrating how to perform agent tasks, ***but without grounded trajectories!***





Control actions

Digital agent learning from online videos

Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

Bowen Baker[*†] bowen@openai.com
Ilge Akkaya[*†] ilge@openai.com
Peter Zhokhov[*†] peterz@openai.com
Joost Huizinga[*†] joost@openai.com
Jie Tang[*†] jietang@openai.com
Adrien Ecoffet[*†] adrien@openai.com
Brandon Houghton[*†] brandon@openai.com
Raul Sampedro[*†] raulsamg@gmail.com
Jeff Clune[*†‡] jclune@gmail.com

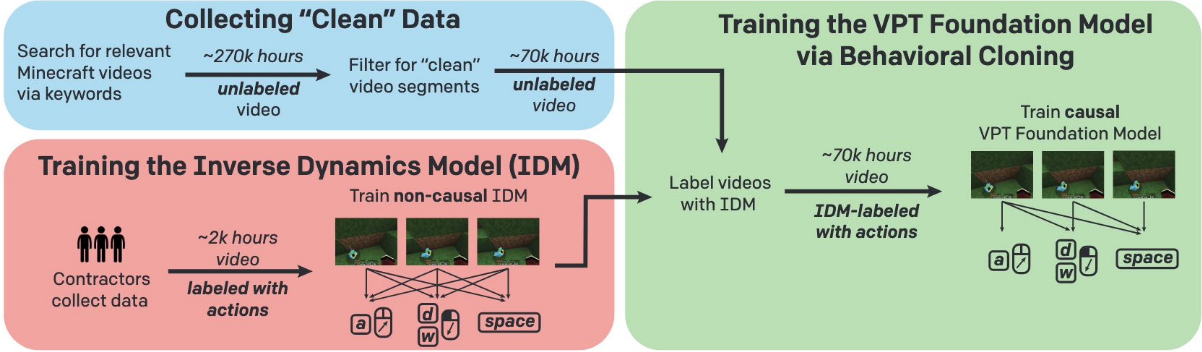


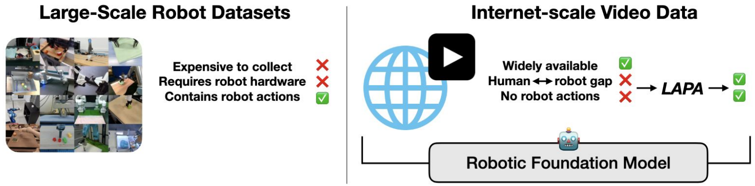
Figure 2: Video Pretraining (VPT) Method Overview.

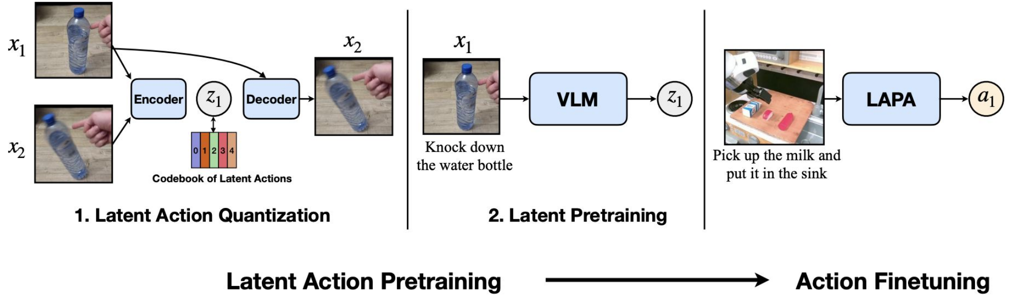
Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos, (Baker et al., 2022)

Robotic learning from internet videos

LAPA: Latent Action Pretraining from Videos

Seonghyeon Ye[*1], Joel Jang[*2],
Byeongguk Jeon[1], Sejune Joo[1], Jianwei Yang[3], Baolin Peng[3], Ajay Mandlekar[4],
Reuben Tan[3], Yu-Wei Chao[4], Yuchen Lin[5], Lars Liden[3],
Kimin Lee[1†], Jianfeng Gao[3†], Luke Zettlemoyer[2†], Dieter Fox[2,4†], Minjoon Seo[1†]
[1]KAIST [2]University of Washington
[3]Microsoft Research [4]NVIDIA [5]Allen Institute for AI
* Equal contribution, † Equal advising





Genie: Generative Interactive Environments, (Genie Team, 2024)
Latent Action Pretraining from Videos, (Ye et al., 2024)
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, (RT-2 Team, 2023)

Internet-scale data is not perfect

Challenges in using internet data for agent training

- Missing grounded action sequences, environmental state info
- Observation-action alignment
- Unclear task objectives from video alone

ImageNet in agent learning?

Agent evaluation

Evaluation in the era of LLMs is hard

Andrej Karpathy
@karpathy

Nice, a serious contender to @lmsysorg in evaluating LLMs has entered the chat.

LLM evals are improving, but not so long ago their state was very bleak, with qualitative experience very often disagreeing with quantitative rankings.

This is because good evals are very difficult to build - at Tesla I probably spent 1/3 of my time on data, 1/3 on evals, and 1/3 on everything else. They have to be comprehensive, representative, of high quality, and measure gradient signal (i.e. not too easy, not too hard), and there are a lot of details to think through and get right before your qualitative and quantitative assessments line up. My goto pointer for some of the fun subtleties is probably the Open LLM Leaderboard MMLU writeup: github.com/huggingface/bl...

Agent evaluation is even more challenging...

Challenges in agent evaluation

- Real-world environmental setup complexity
- Task coverage
- Open-ended success criteria
  - Multiple valid solution paths
  - Cannot script evaluation metrics, need for human judgment
- Evaluation beyond task success

Agent evaluation

- via benchmarks
- via LLMs/VLMs
- via crowdsourcing

Agent evaluation

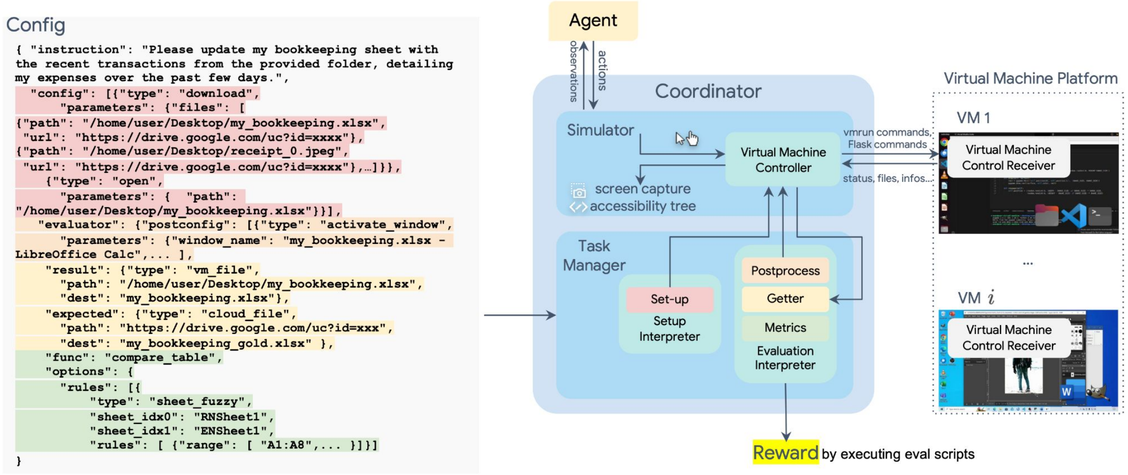
via benchmarks

How to define good agent benchmarks?

- Natural and challenging tasks

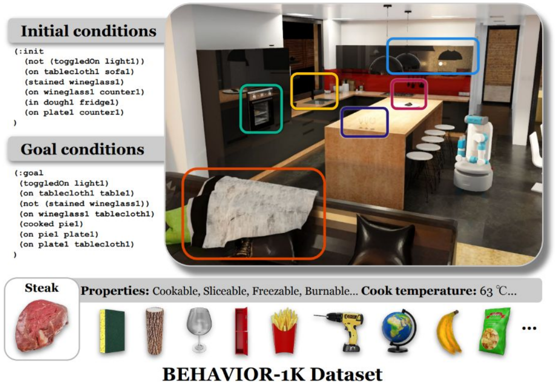
How to define good agent benchmarks?

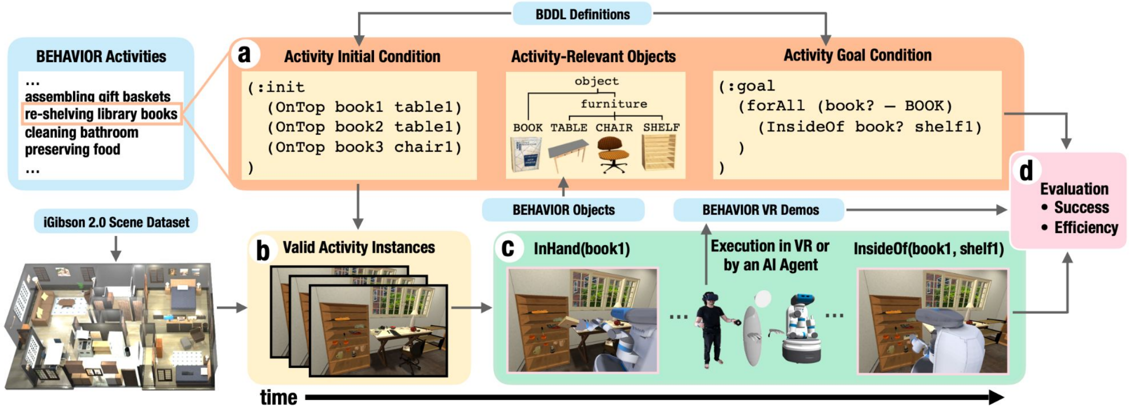
- Natural and challenging tasks
- Good agent evaluation framework
  - Realistic agent environment
  - Automatic initial task state setup
  - Automatic task evaluation: execution-based scripts to compare final states



OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments, (Xie et al., 2024)
Browsergym: a Gym Environment for Web Task Automation (Drouin et al., 2024)

Robotic task evaluation





BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments, (Srivastava et al., 2021)

More agent evaluation metrics

Other agent evaluation metrics

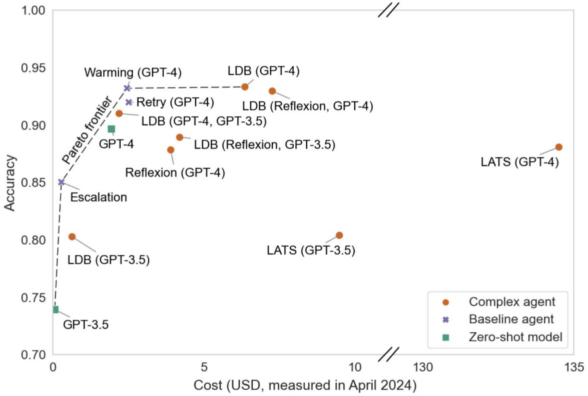
- Latency - efficiency
  - Compute aware success rate
  - Real time evaluation
- Robustness
  - Generalization to unseen domains, tasks, apps

AI Agents That Matter

Sayash Kapoor*, Benedikt Stroebl*, Zachary S. Siegel, Nitya Nadgir, Arvind Narayanan

Princeton University

July 2, 2024



AI Agents That Matter, (Kapoor et al., 2024)

More agent evaluation metrics

Other agent evaluation metrics

- Safety - will be covered by Diyi

Jailbreaking LLM-Controlled Robots, (Robey, et al., 2024)

Limitations of agent benchmarks

- Only can write evaluation scripts for very limited tasks, time-consuming
- Cannot script evaluation metrics for open-answer tasks

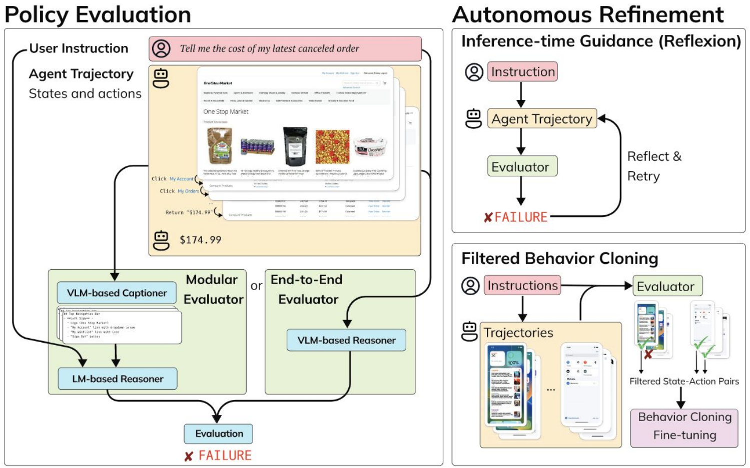
Possible solution: leveraging LLM/VLMs to automatically evaluate agent tasks?

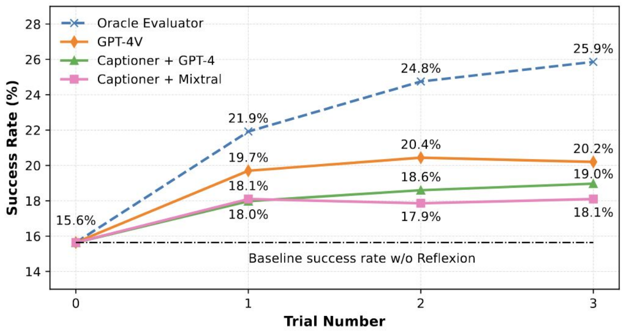
Agent evaluation

via LLMs/VLMs

Automatic agent evaluation

Automatically evaluate user instructions and arbitrary agent trajectories with LLM/VLMs

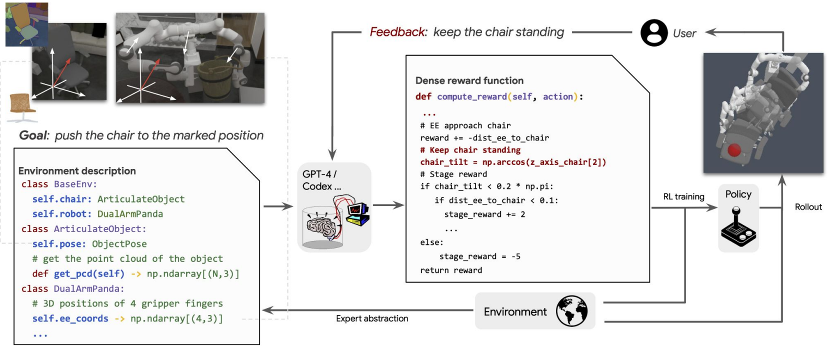


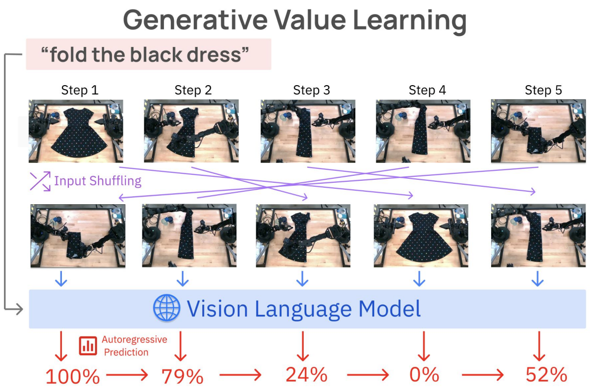


Autonomous Evaluation and Refinement of Digital Agents, (Pan et al., 2024)

Automatic agent evaluation

- Leveraging coding ability of LLMs to automatically generate reward functions
- Leveraging the world knowledge embedded in VLMs to evaluate task progress





Text2Reward: Reward Shaping with Language Models for Reinforcement Learning, (Xie et al., 2023)
Vision Language Models are In-Context Value Learners, (Ma et al., 2024)

Limitations of automatic agent evaluation

- Limited foundation model capabilities
- Missing personalized task evaluation

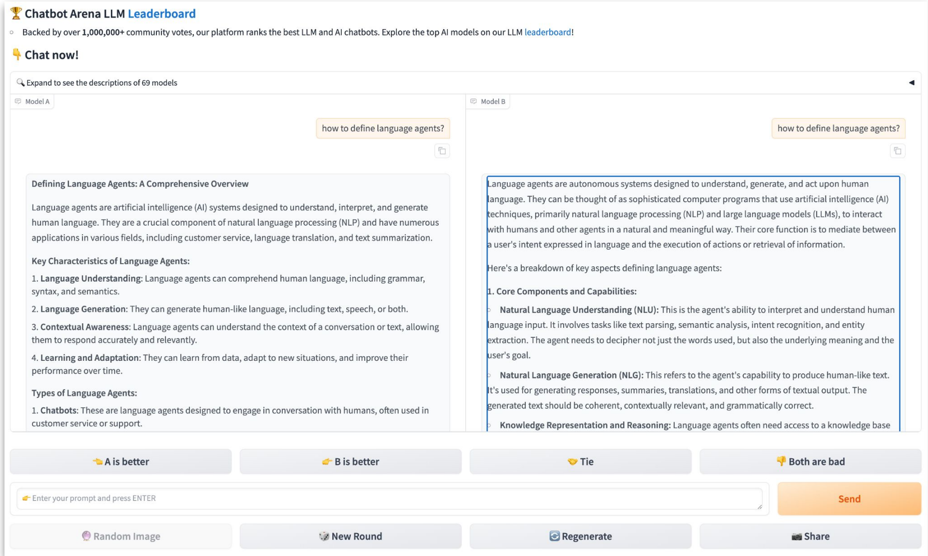
Possible solution: how about evaluating agent tasks via crowdsourcing from real users?

- Personalized and robust success criteria capture
- Diverse task scenarios and environments
- Natural interaction and feedback loops
- Hard to overfit

Agent evaluation
via crowdsourcing

Chatbot arena

Chatbot Arena is not embodied for agent evaluation

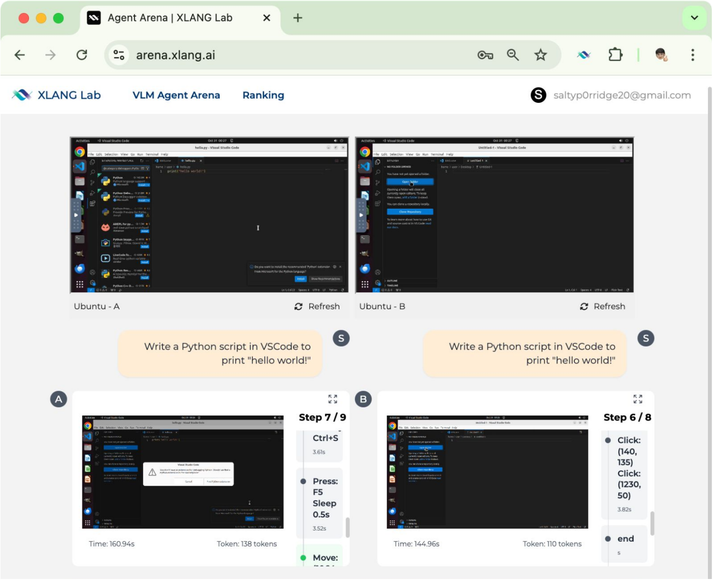


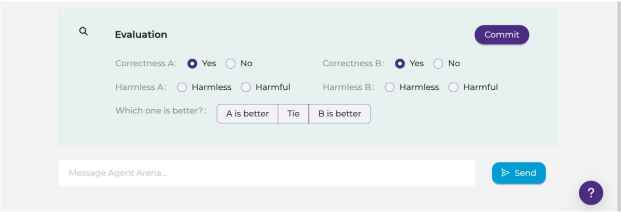
Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. (Chiang et al., 2024)

Agent arena for digital agent task evaluation

Computer Agent Arena: https://arena.xlang.ai

- an open evaluation platform where users can compare LLM/VLM-based AI agents performing real-world computer tasks, ranging from general computer use to specialized workflows like coding, data analysis, and video editing





Current agent arena leaderboard

Computer Agent Arena: https://arena.xlang.ai

- More advanced AI agents are expected to emerge in the near future!

XLANG Lab Computer Agent Arena Leaderboard Login

Computer Agent Arena Leaderboard by XLANG Lab

Computer Agent Arena by XLANG Lab is an open evaluation platform where users can compare LLM/VLM-based AI agents performing real-world computer tasks, ranging from general computer use to specialized workflows like coding, data analysis, and video editing.

Ranking	VLM	Agent	Correct Rate	Win Rate	Evaluate Times	Overall
1	claude-3-5-sonnet-20241022	vision_based	0.62	0.77	56	1165.75
2	claude-3-5-sonnet-20241022	claude_computer_use	0.44	0.55	52	1073.60
3	claude-3-5-sonnet-20240620	vision_based	0.56	0.60	34	1066.10
4	gpt-4o-2024-08-06	vision_based	0.52	0.54	61	1000.31
5	claude-3-opus-20240229	vision_based	0.13	0.43	15	992.06
6	gpt-4-turbo-2024-04-09	vision_based	0.29	0.32	38	893.84
7	gpt-4o-mini-2024-07-18	vision_based	0.12	0.27	67	830.81