Computer Use Agents

JY Koh

11-768: AI Agents

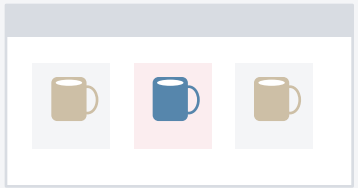
15 Sep 2026

Carnegie Mellon University

Computer Use Agents (CUAs)

CUAs observe rendered GUIs, execute actions, and checks what changed.

BROWSER

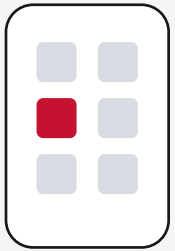


Shop across pages

DESKTOP

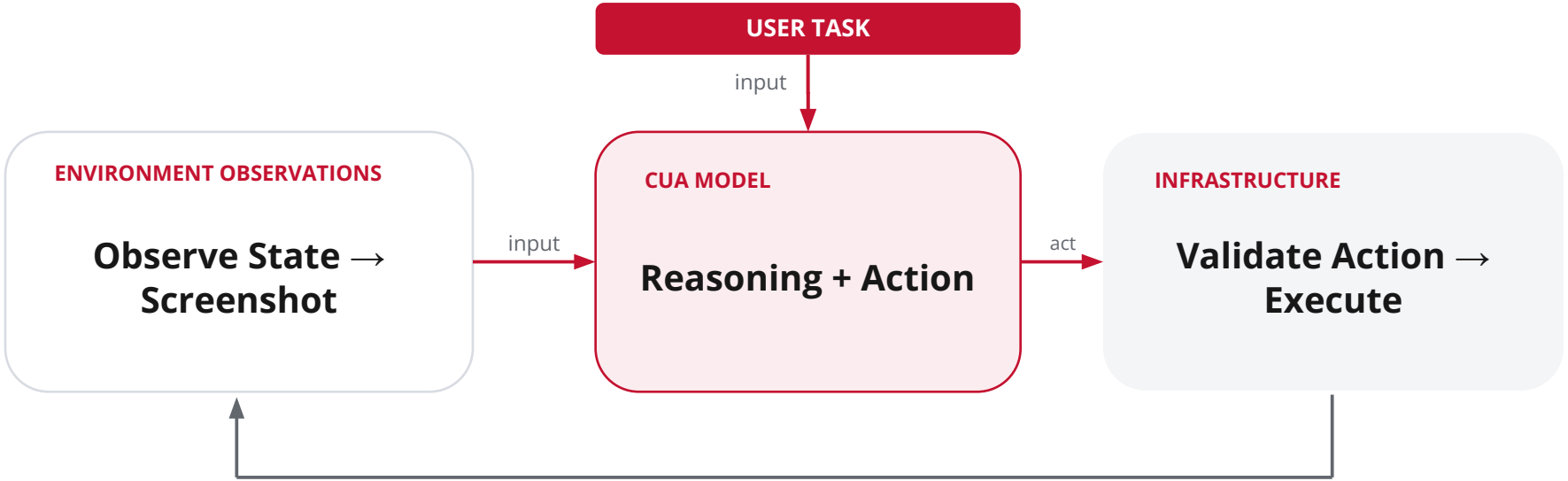
Edit files across apps

MOBILE



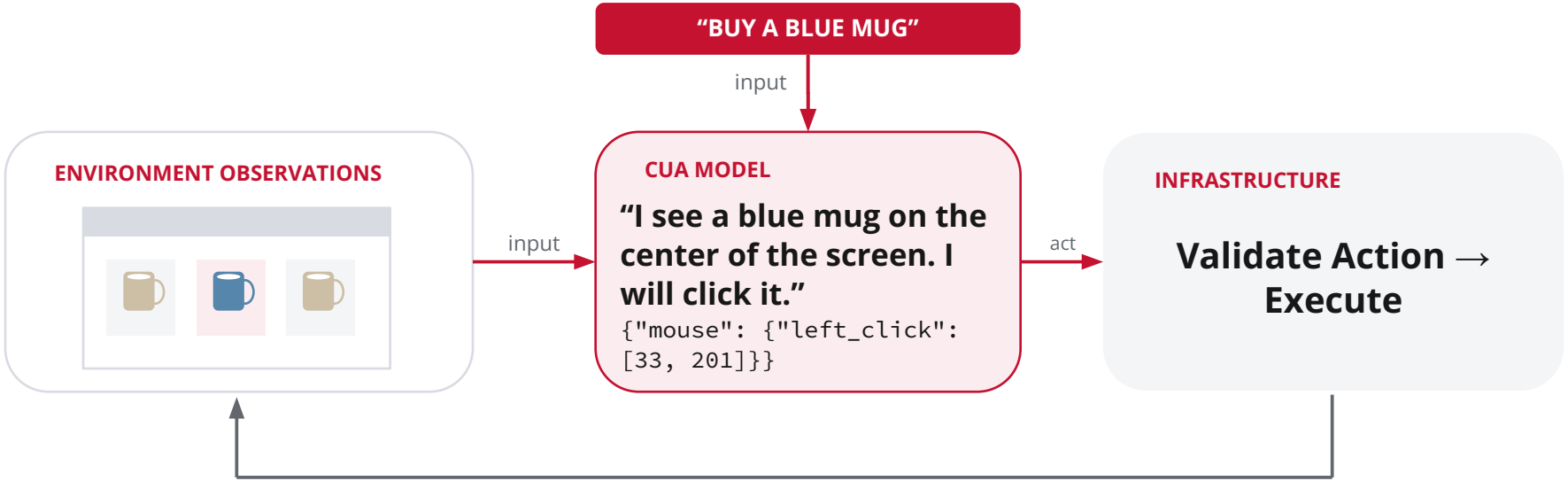
Navigate screens and gestures

Observe, Reason, Act

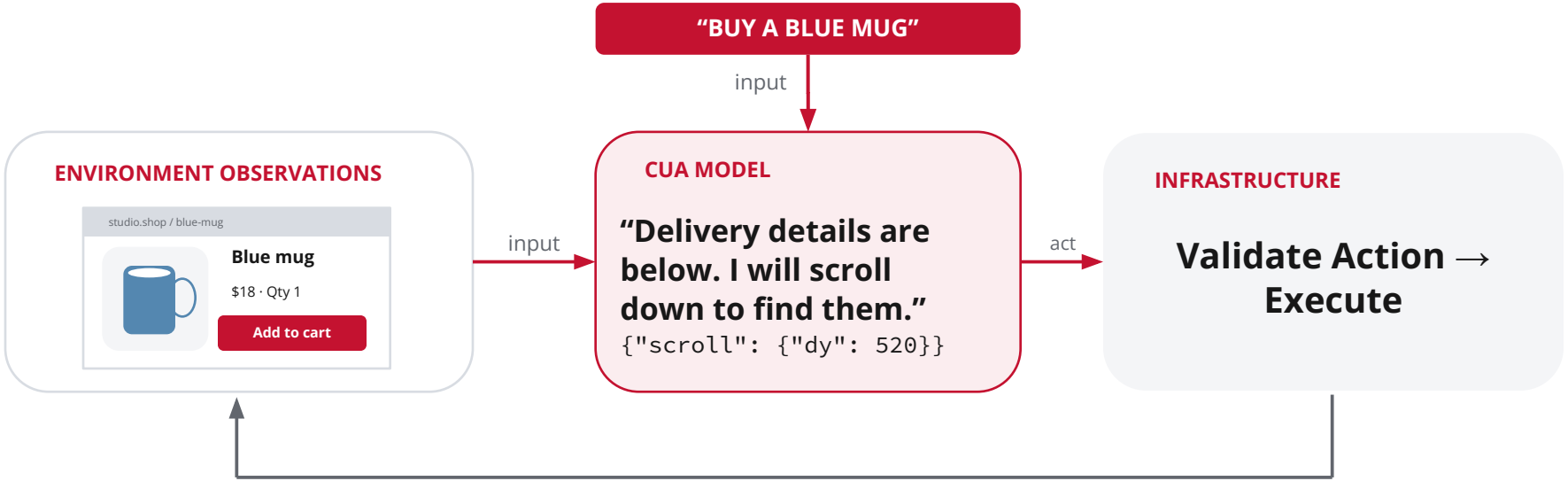


Actions change the state, a new screenshot observation is provided to the model.

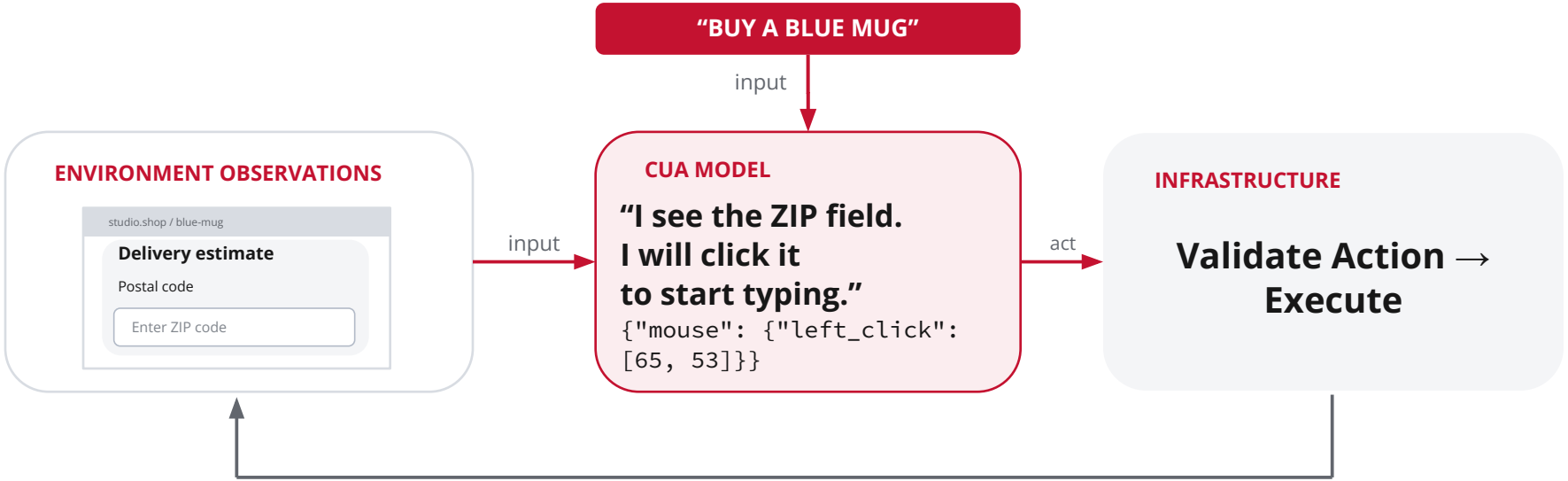
Observe, Reason, Act



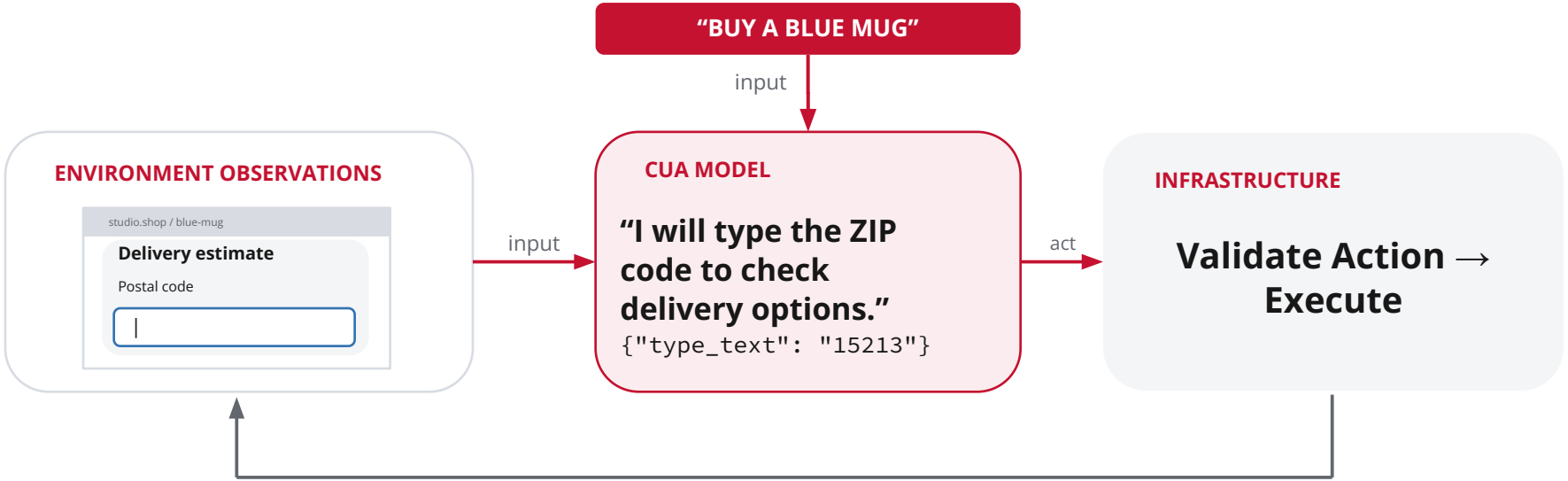
Observe, Reason, Act



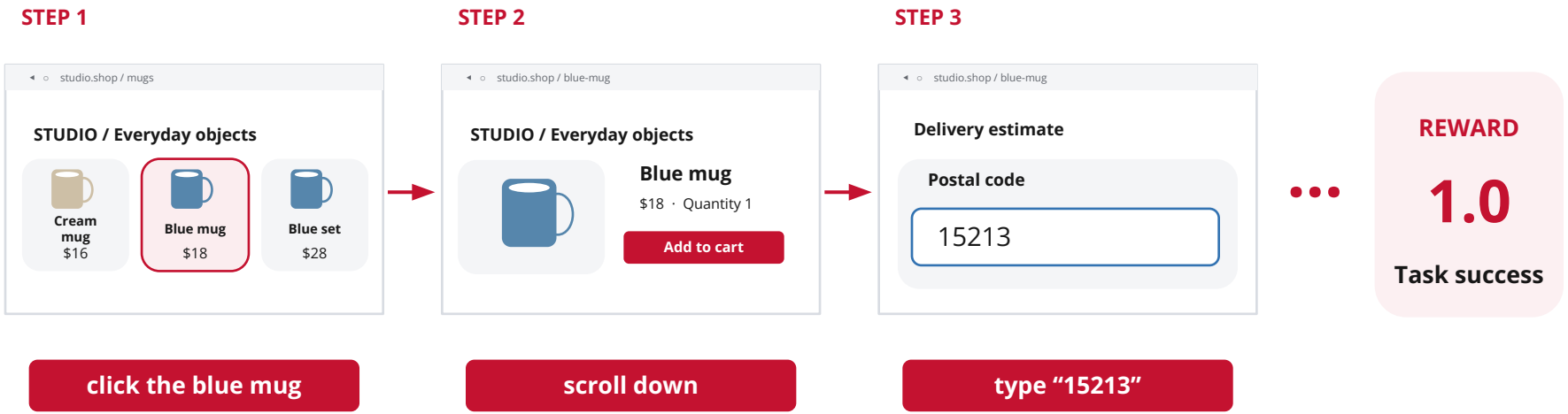
Observe, Reason, Act



Observe, Reason, Act



Observe, Reason, Act



The loop continues until task completion, then a reward is computed.

A Brief History of CUAs: Toy Environments

Learning to use greatly simplified toy web interfaces

2017–2022 | 2022–2024 | 2025–2026 | Sep 2026

MiniWoB: Shi et al., ICML 2017 · MiniWoB++: Liu et al., ICLR 2018 · WebShop: Yao et al., NeurIPS 2022

A Brief History of CUAs: Toy Environments

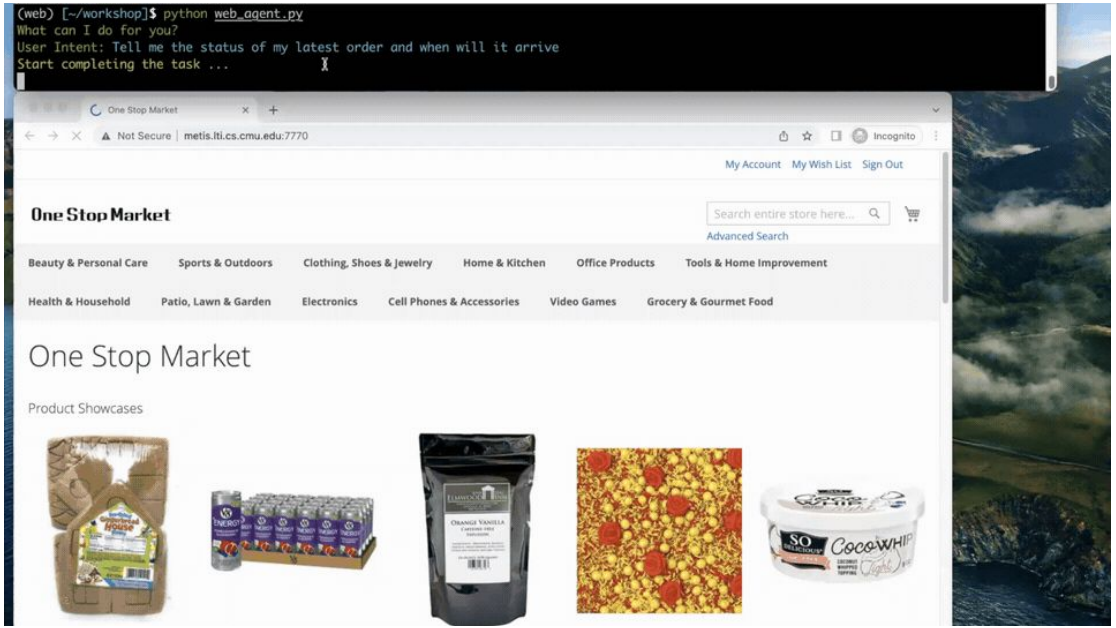
Shopping with real products and language goals

2017–2022 | 2022–2024 | 2025–2026 | Sep 2026

WebShop: Yao et al., NeurIPS 2022 · Official environment demo

A Brief History of CUAs: Realistic Environments

From toy tasks to realistic web interfaces

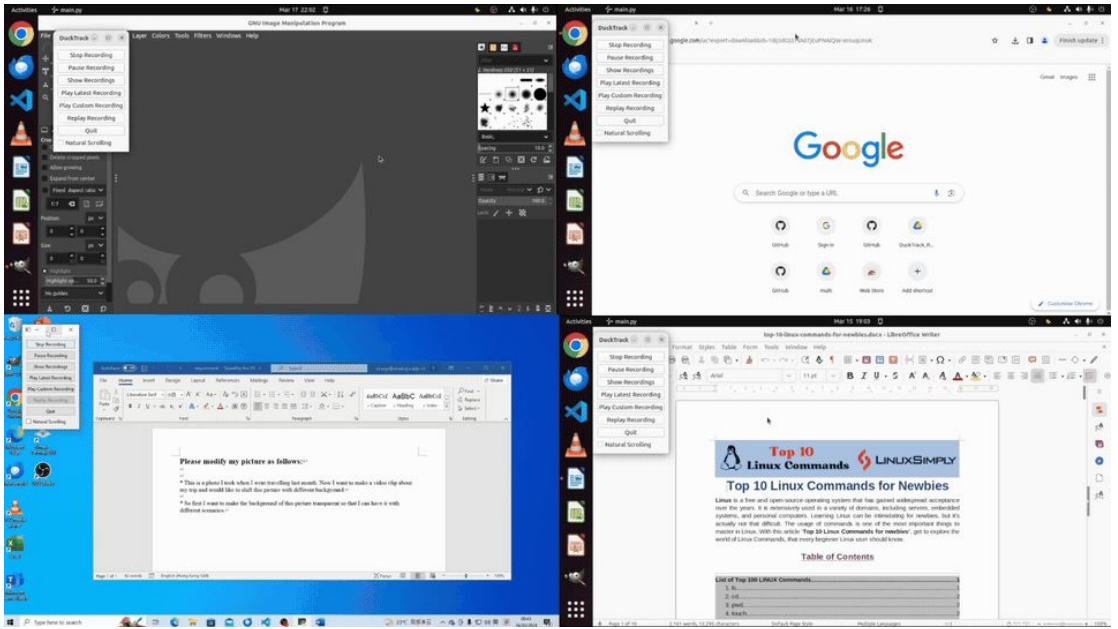


2017–2022 | **2022–2024** | 2025–2026 | Sep 2026

WebArena: Zhou et al., ICLR 2024 · VisualWebArena: Koh et al., ACL 2024

A Brief History of CUAs: Realistic Environments

Full (realistic) desktop workflows



2017–2022 | **2022–2024** | 2025–2026 | Sep 2026

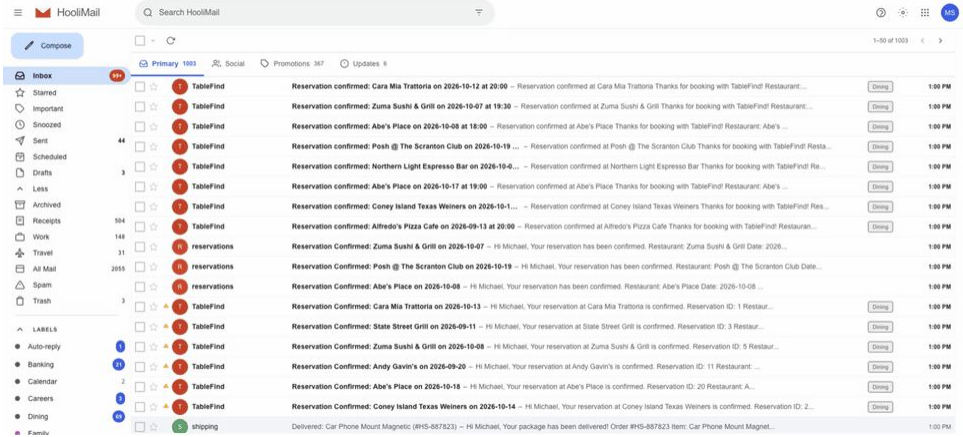
OSWorld: Xie et al., NeurIPS 2024 · Official overview video

A Brief History of CUAs: Realistic Envs and Data

MYPCBENCH

Realistic apps with personal history

Explore mail, calendar events, and project issues in one shared world.



2017–2022 | 2022–2024 | **2025–2026** | Sep 2026

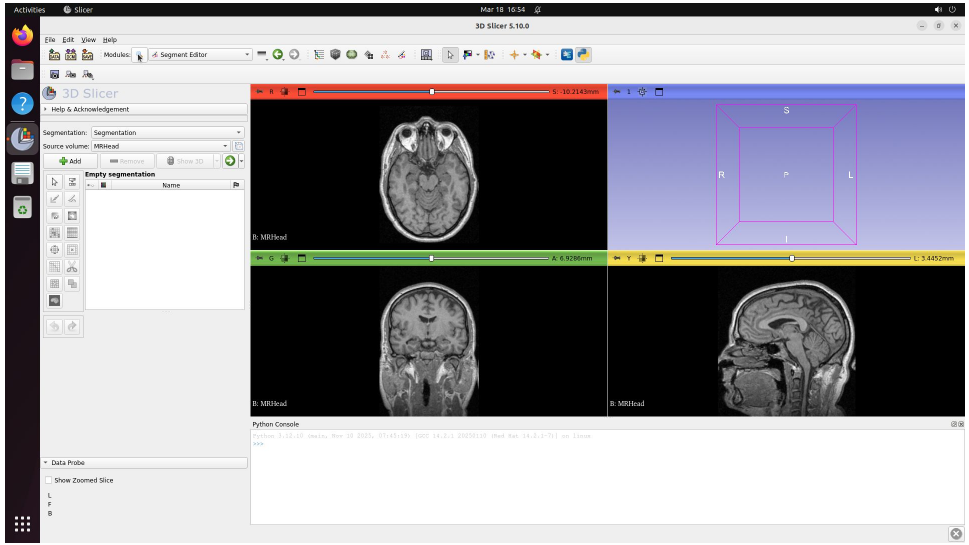
Jang et al., 2026 · MyPCBench · Recorded environment walkthrough (33 s)

A Brief History of CUAs: Realistic Envs and Data

CUA-WORLD

Real software, realistic tasks

Specialized desktop software with loaded data and realistic tasks.



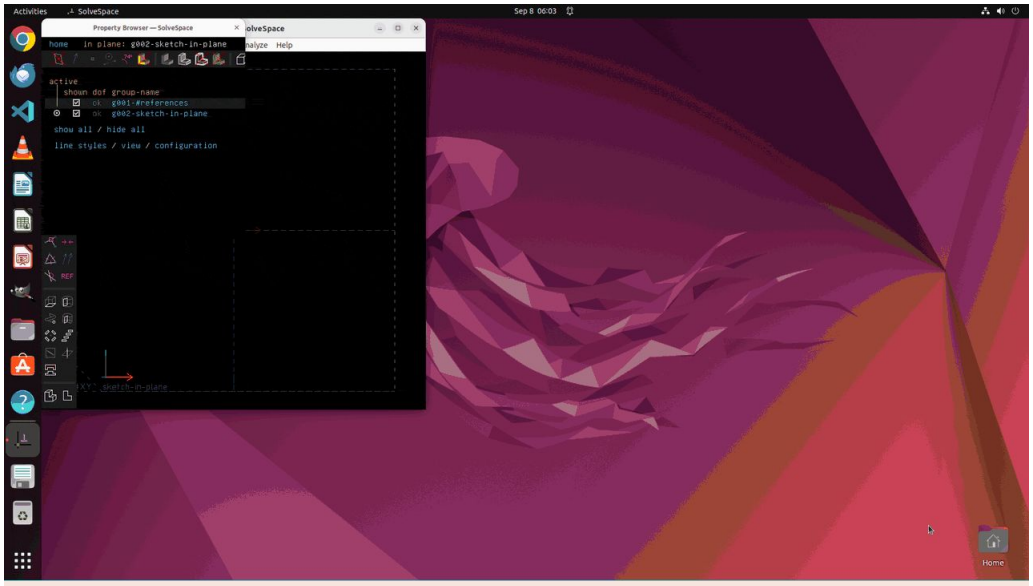
2017–2022 | 2022–2024 | **2025–2026** | Sep 2026

Aggarwal et al., 2026 · Gym-Anything / CUA-World

A Brief History of CUAs: Real World Use

GPT-6 ASTRA

Very capable and fast, still (very) expensive.



2017–2022 | 2022–2024 | 2025–2026 | **Sep 2026**

Koh, Sep 2026 · What's the Point of Computer Use Agents? · OSWorld 2.0 SolveSpace demo (30×)

Benchmarks & Evaluation

What should count as success?

The same task can be evaluated in multiple different ways.

ACTION

Did it click the target?

Fast grounding / imitation tests

OUTCOME

Is the right item in the cart?

State checks or evidence judging

PROCESS

Did it respect the constraints?

No checkout; no unwanted changes

Evaluators should match the capability you want to measure.

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

Static evals: Was this action correct?

TASK: Add a blue mug under $20 to the cart. Do not purchase.

SCREENSHOT

REFERENCE RULE

$(x, y) \in$ target box

For recorded actions:
Match target, operation, and input value.

STEP SCORE

1.0

correct point / step

No environment rollout

A correct step is not a completed task.

BENCHMARK EXAMPLES

ScreenSpot-Pro · Li et al., 2025

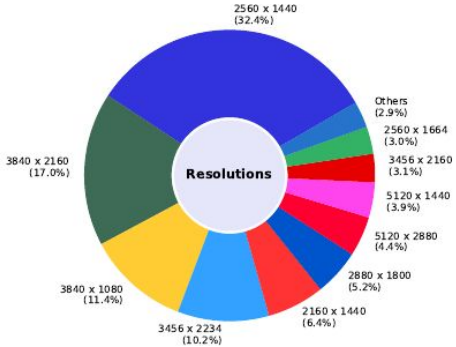
AITW · Rawles et al., 2023

Mind2Web · Deng et al., 2023

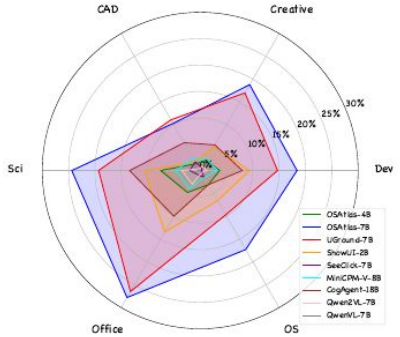
AndroidControl · Li et al., 2024

ScreenSpot-Pro: Grounding in professional UIs

(a) Software categories and the number of tasks.



(b) Resolution distribution.



(c) Results of representative GUI Grounding models across 6 categories of applications.

BENCHMARK STATS

1,581
screenshots

23
applications

3
operating systems

EVALUATION CRITERION

Point / region matching

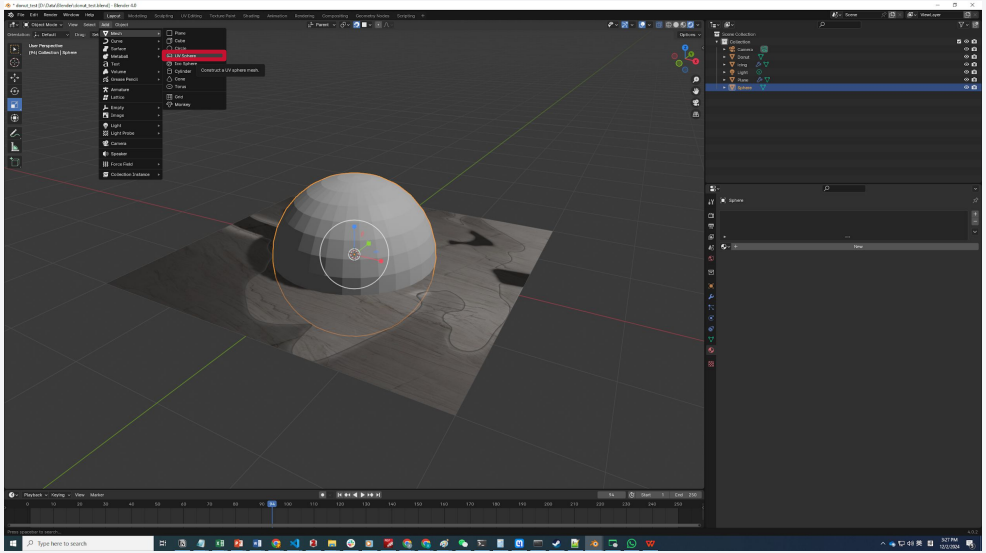
High-resolution professional UIs make grounding a benchmark of its own.

Source: Li et al., 2025, Fig. 1

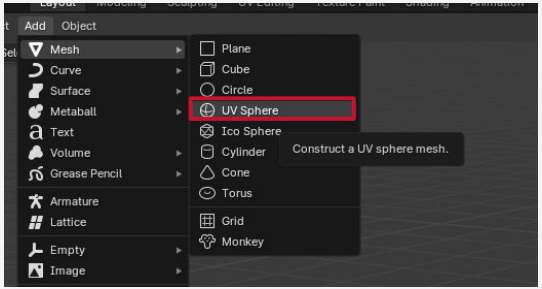
ScreenSpot-Pro: Grounding in professional UIs

TASK: Construct a UV sphere mesh.

SCREENSHOT / BLENDER



GROUND TRUTH

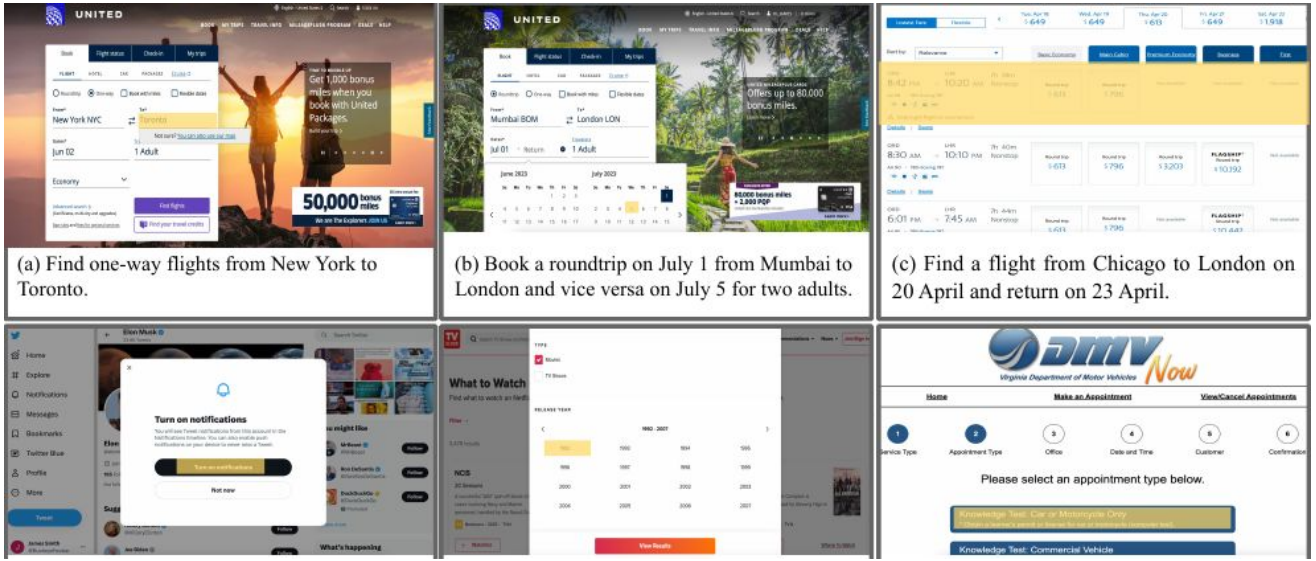


Box in original pixels
(417, 132)–(576, 152)

Point inside = 1
Original: 2560 × 1440

Li et al., 2025 · ScreenSpot-Pro, blender_windows_1

Mind2Web: Offline generalization across the web



BENCHMARK STATS

2,350
tasks

137
websites

31
domains

EVALUATION CRITERION

Offline action matching

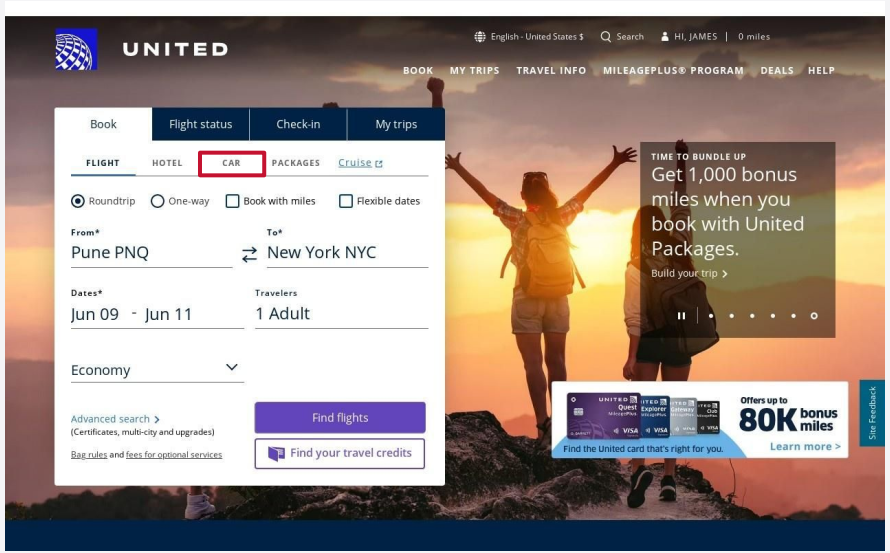
Next-action imitation is not the same thing as task completion.

Source: Deng et al., NeurIPS 2023, Fig. 1

Mind2Web: Offline generalization across the web

TASK: Rent a car in Brooklyn–Central, NY, for April 9–15.

RECORDED SCREENSHOT / UNITED



Ground-truth target outlined in red

GROUND-TRUTH ACTION

CLICK
Target: CAR tab
Value: none

Element: bookCarTab

Match the target element and operation. No task rollout is performed.

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

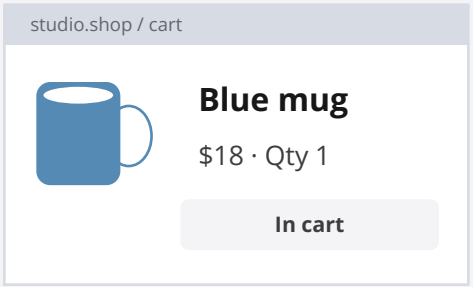
Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

Programmatic state: Test the resulting world

TASK: Add a blue mug under $20 to the cart. Do not purchase.

FINAL APP STATE



EXECUTABLE CHECKS

```
cart.item == blue_mug
cart.price < 20
orders.count == 0

assert all(checks)
```

TASK REWARD

PASS

reward = 1.0

Database / file / app state

Repeatable and precise—but only for conditions the tests cover.

BENCHMARK EXAMPLES

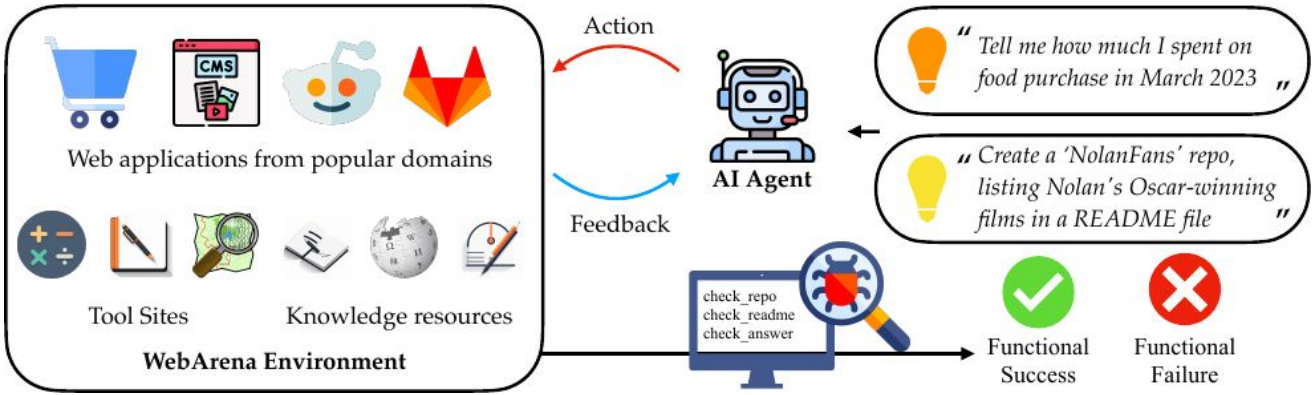
WebArena · Zhou et al., 2024

AndroidWorld · Rawles et al., 2024

OSWorld · Xie et al., 2024

WorkArena · Drouin et al., 2024

WebArena: Reproducible sites, executable goals



BENCHMARK STATS

812
long-horizon tasks

4
self-hosted domains

functional
success checks

EVALUATION CRITERION

Programmatic end state

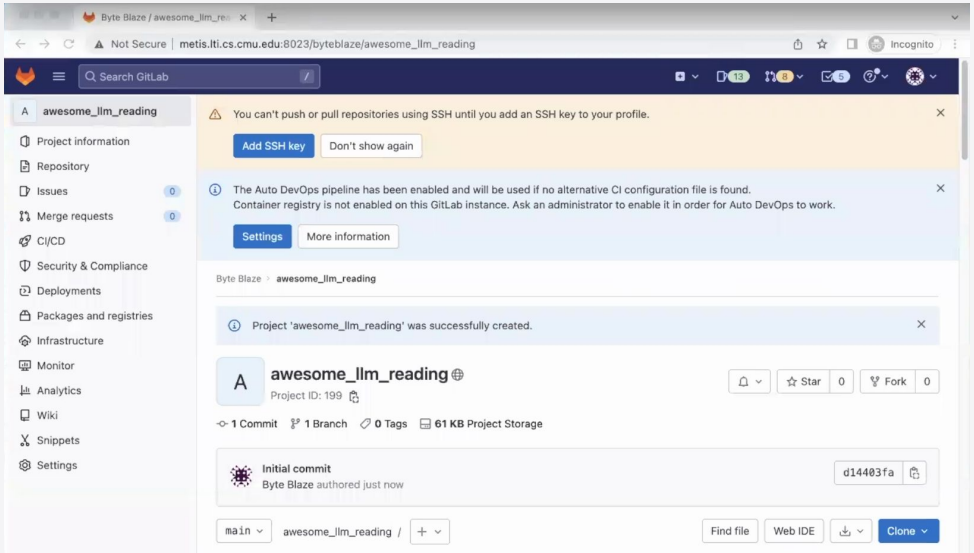
Evaluators check environment state, not whether the trace "looks right".

Source: Zhou et al., ICLR 2024, Fig. 1

WebArena: Reproducible sites, executable goals

TASK: Create an empty repository named awesome_llm_reading.

OFFICIAL DEMO / CREATED PROJECT



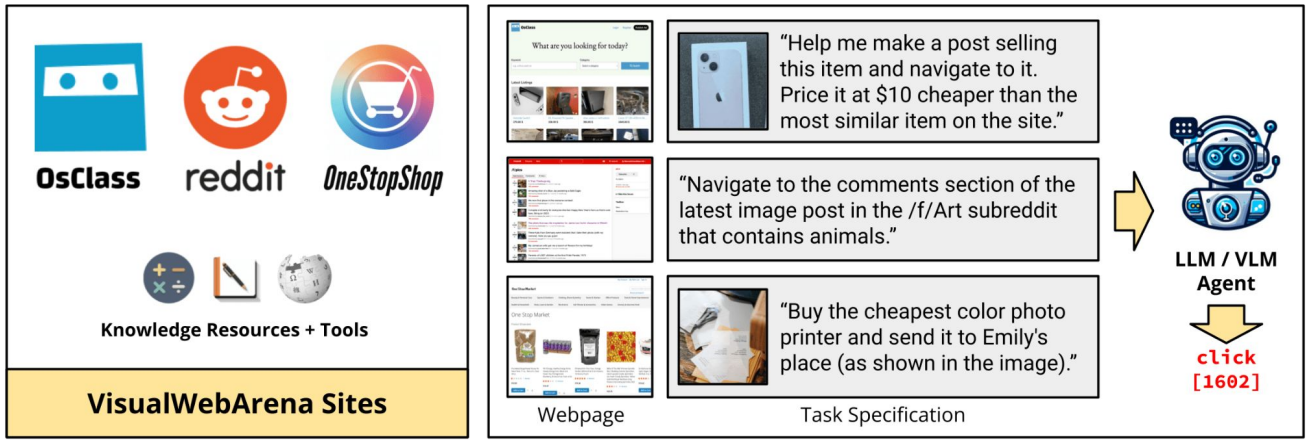
PROGRAMMATIC VERIFIER

```
name = task.project_name
p = "/byteblaze/"
page.goto(GITLAB+p+name)
html = page.content()
reward = int(
  name in html.lower()
)
```

Task 476 (simplified).

Checks name at target URL.
Emptiness is not checked.

VisualWebArena: Vision is part of the task



BENCHMARK STATS

910
tasks

3
self-hosted sites

25.2%
image-input tasks

EVALUATION CRITERION

Multimodal + end state

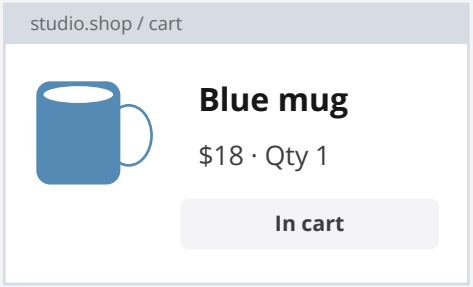
Visual understanding changes which tasks are even expressible.

Source: Koh et al., ACL 2024, Fig. 1

LLM / VLM judging: Evaluate the evidence

TASK: Add a blue mug under $20 to the cart. Do not purchase.

OBSERVED EVIDENCE



Screenshots + action history

LLM / VLM + RUBRIC

✓ Blue mug present
✓ Price below $20
✓ No purchase shown

Rubric → evidence → verdict

JUDGMENT

PASS

"The visible cart meets the requested criteria."

Flexible evidence judging needs calibration against human labels.

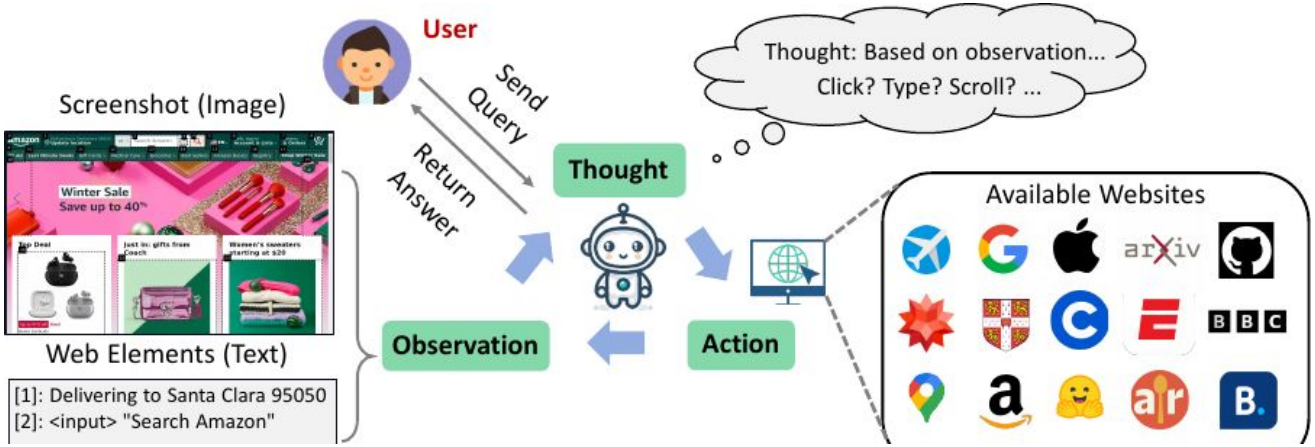
BENCHMARK EXAMPLES

WebVoyager · He et al., 2024

Online-Mind2Web / WebJudge · Xue et al., 2025

CUA-World-Long · Aggarwal et al., 2026

WebVoyager: Live web, open-ended outcomes



BENCHMARK STATS

643
tasks

15
live websites

85.3%
judge-human agreement

EVALUATION CRITERION

VLM-as-judge

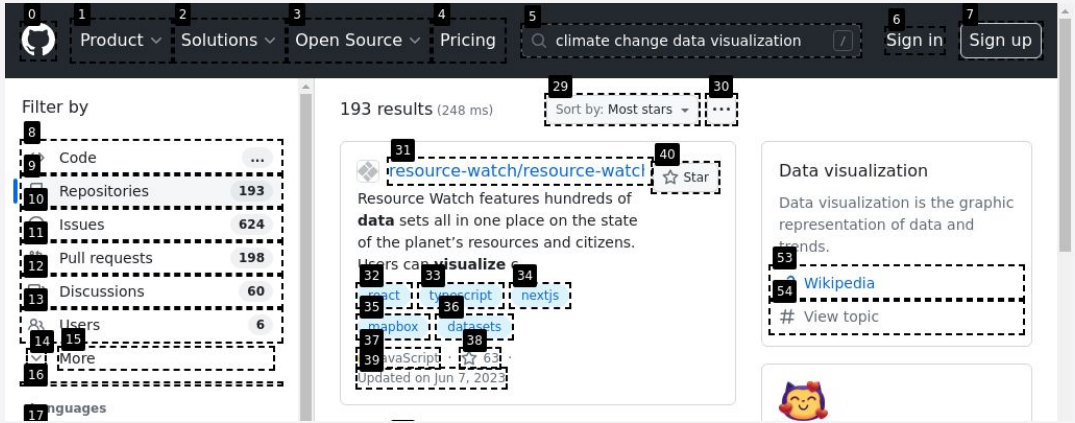
Realism rises; reproducibility falls.

Source: He et al., ACL 2024, Fig. 1

WebVoyager: Live web, open-ended outcomes

TASK: Find the most-starred GitHub project for climate-change data visualization.

PUBLISHED TRACE / GITHUB SEARCH



Historical trace, not current GitHub rankings

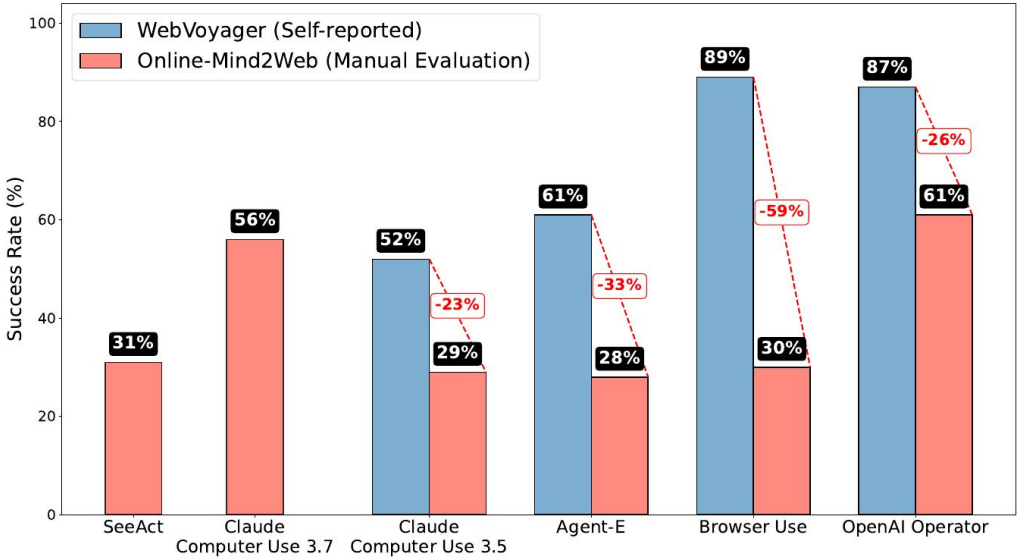
VLM JUDGE INPUTS

Task + final answer + last k screenshots

Recorded answer resource-watch/ resource-watch: 63 stars

Judge checks the evidence and returns **SUCCESS** or **NOT SUCCESS**.

Online-Mind2Web: Live sites expose benchmark drift



BENCHMARK STATS

300
tasks

136
live websites

85.7%
WebJudge-human agreement

EVALUATION CRITERION

Human + LLM judge

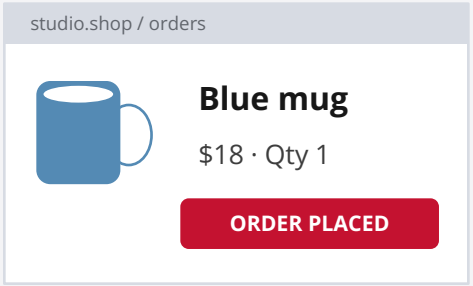
Live benchmarks drift—and judge quality becomes part of the score.

Source: Xue et al., COLM 2025, Fig. 1 (cropped)

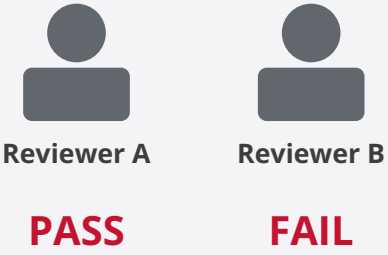
Human review: Resolve ambiguous outcomes

TASK: Add a blue mug under $20 to the cart. Do not purchase.

REVIEW EVIDENCE



INDEPENDENT REVIEW



ADJUDICATION

FAIL

Order confirmation shows a forbidden purchase.

Resolve disagreement using evidence

Human review can score tasks or audit an automatic evaluator.

BENCHMARK EXAMPLES

WebVoyager · human scoring & judge audit · He et al., 2024

Online-Mind2Web · human labels & review · Xue et al., 2025

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

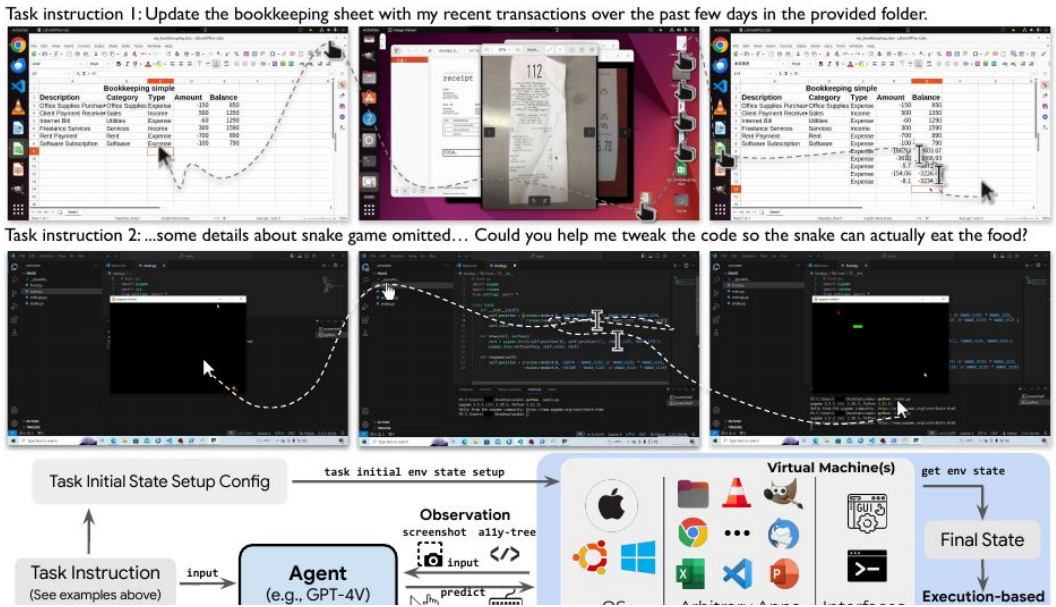
02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

OSWorld: Real desktops, custom end-state checks



BENCHMARK STATS

369
tasks

9
applications

custom
task evaluators

EVALUATION CRITERION

Custom end-state checks

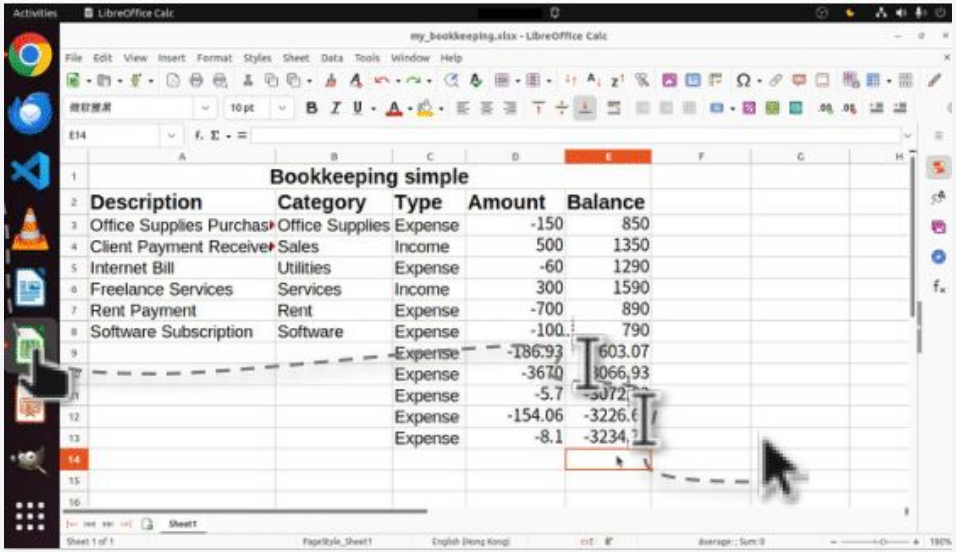
Every task ships with setup logic and application-specific state checks.

Source: Xie et al., NeurIPS 2024, Fig. 1

OSWorld: Real desktops, custom end-state checks

TASK: Update the bookkeeping sheet using the receipts in the provided folder.

PAPER EXAMPLE / UPDATED WORKBOOK



REFERENCE FILE + RULES

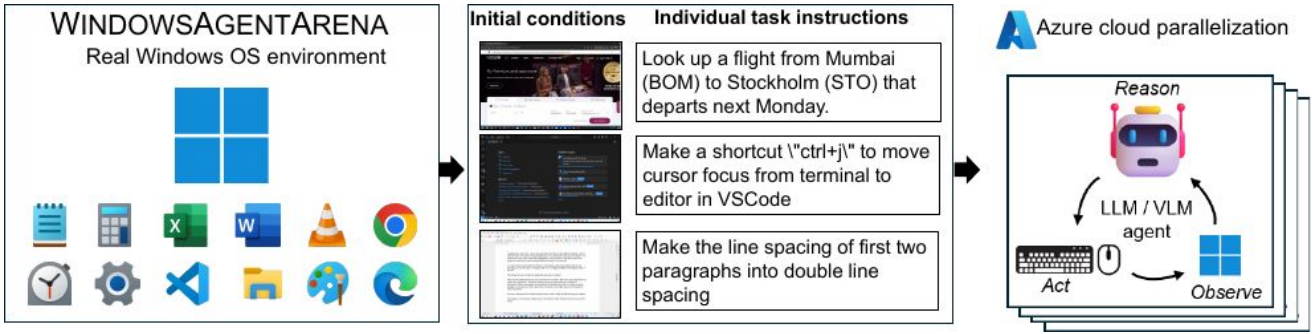
```
compare_table(
  saved_workbook,
  gold_workbook,
  rules
)
```

Reference checks
A1:E8 unchanged
C9:C13 expense labels
Amounts + running balances

D9 ≈ −186.93
E9 ≈ 603.07

Xie et al., NeurIPS 2024 · Fig. 1 + bookkeeping task 8e116af7

WindowsAgentArena: Windows at cloud scale



BENCHMARK STATS

154
tasks

11
programs

20 min
parallel sweep

EVALUATION CRITERION

Programmatic end state

Parallel VMs turn multi-day desktop evaluations into a cloud-scale sweep.

Source: Bonatti et al., 2024, Fig. 1

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

Hybrid / trajectory: Check outcome and process

TASK: Add a blue mug under $20 to the cart. Do not purchase.

ACTION TRACE

```
1  click(mug)
2  click(add_to_cart)
3  click(place_order)
```

TWO KINDS OF CHECK

STATE CHECK

✓ Correct item

PROCESS CHECK

✕ Purchase was made

FINAL VERDICT

FAIL

Outcome-only grading could miss the violation.

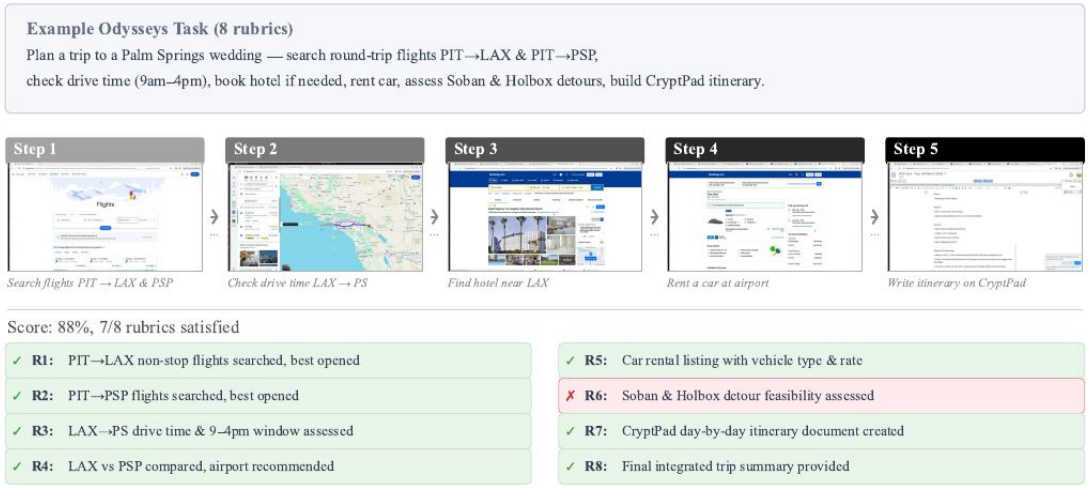
A plausible deliverable is not enough if the process violated the task.

BENCHMARK EXAMPLES

WeaveBench · trajectory-aware judge · Li et al., 2026

OSWorld 2.0 · completion + separate safety audit, 2026

Odysseys: Long-horizon tasks on the live web



BENCHMARK STATS

200
live-web tasks

6.1
rubric items per task (avg.)

100
steps · main evaluation cap

EVALUATION CRITERION

Rubric-based LLM judge

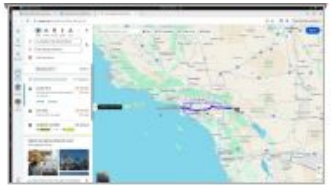
Grade progress across the workflow—not just a single final screenshot.

Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks. Jang*, Koh*, Fried & Salakhutdinov, 2026.

Odysseys: Score progress across the whole trip

TASK: Plan a Palm Springs wedding trip, compare airports, and build an itinerary.

Flights



Drive window

Hotel if needed

Car rental

CryptPad itinerary

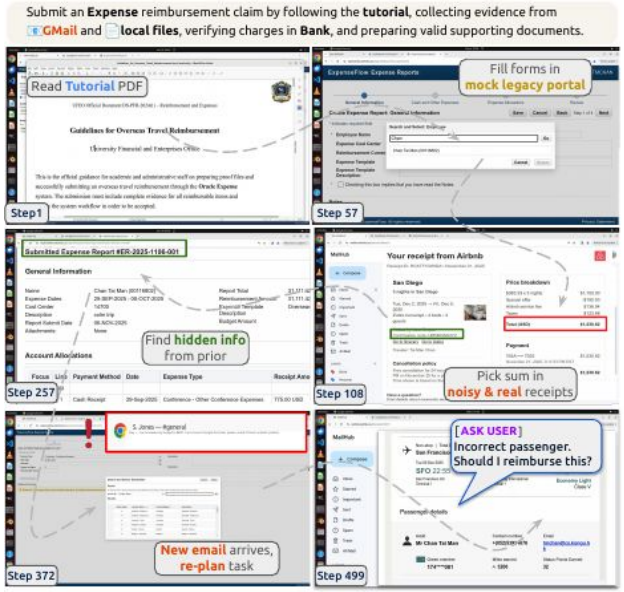
RUBRIC CHECKPOINTS / PUBLISHED ILLUSTRATION

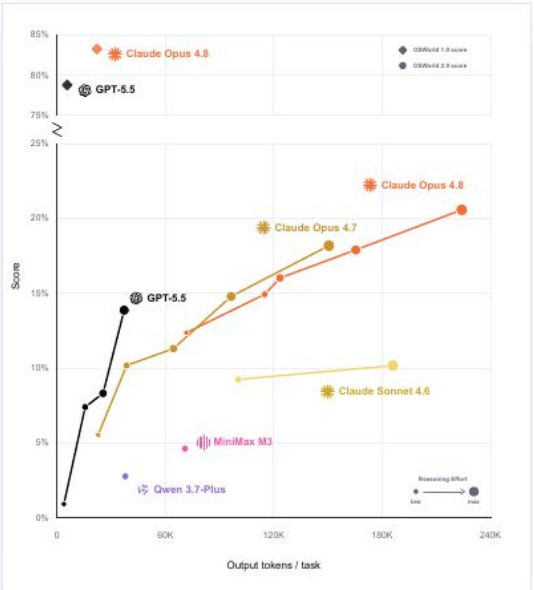
R1	Compare PIT–LAX flights	✓	R5	Open a priced car-rental option	✓
R2	Compare PIT–PSP flights	✓	R6	Evaluate both restaurant detours	×
R3	Check the 9am–4pm drive window	✓	R7	Create an editable daily itinerary	✓
R4	Recommend LAX or PSP	✓	R8	Summarize the complete trip plan	✓

7 / 8 satisfied = 0.875 average **Perfect completion = 0.0**

Jang et al., 2026 · Published trip-planning illustration; rubric text condensed

OSWorld 2.0: Hours-long workflows





BENCHMARK STATS

108
workflows

1.6 h
median human time

318
average agent calls

EVALUATION CRITERION

Partial + binary checks

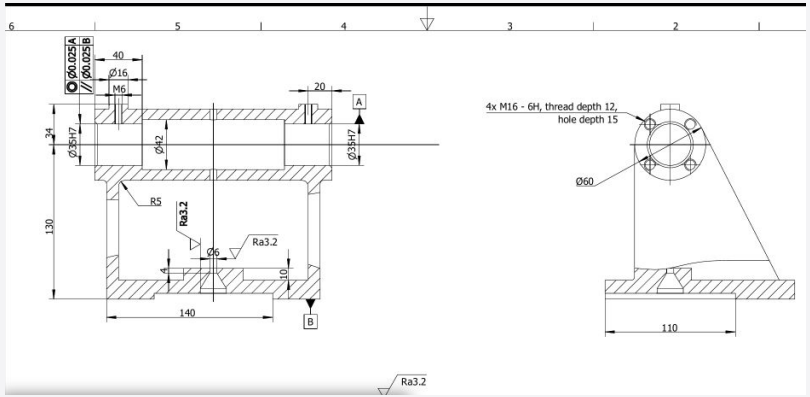
The frontier shifts from clicks to hidden state, changing requirements, and verification.

Source: Yuan, Zhou, Xiong et al., 2026, Fig. 1

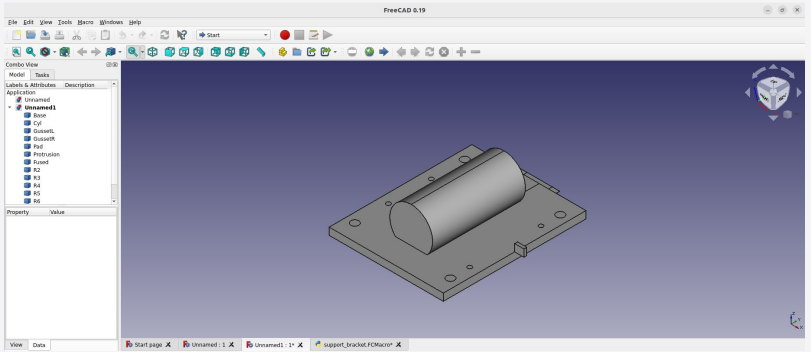
OSWorld 2.0: A file can exist and still be wrong

TASK: Build a FreeCAD support bracket and export a STEP model matching the drawing.

REFERENCE DRAWING / STEP 27



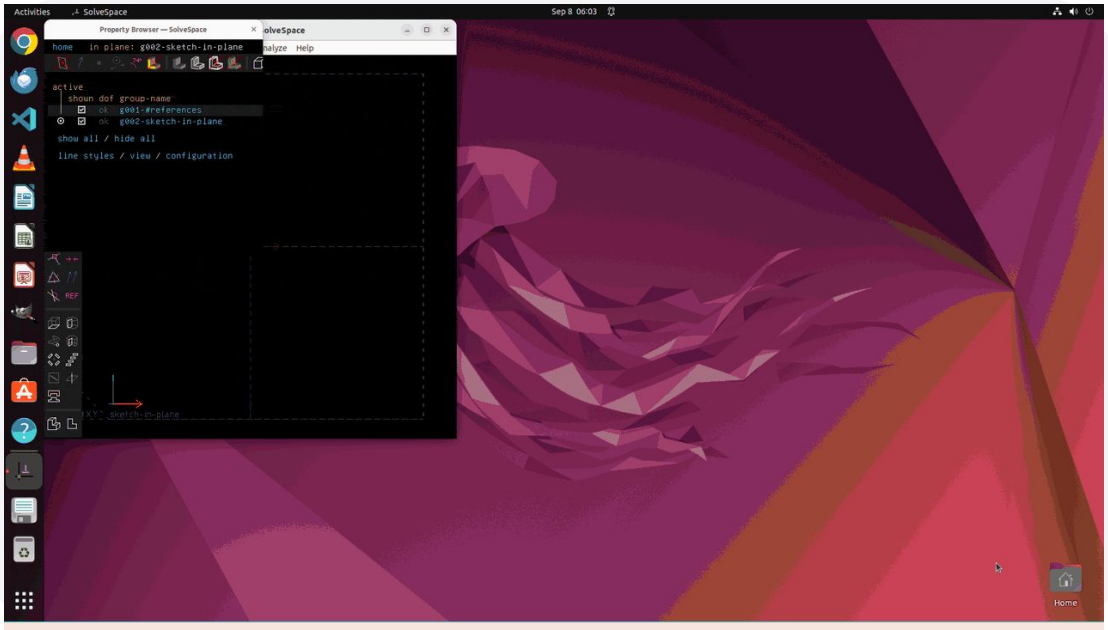
EXPORTED MODEL / STEP 251



Yuan, Zhou, Xiong et al., 2026 · Official Task103 failure case (cropped)

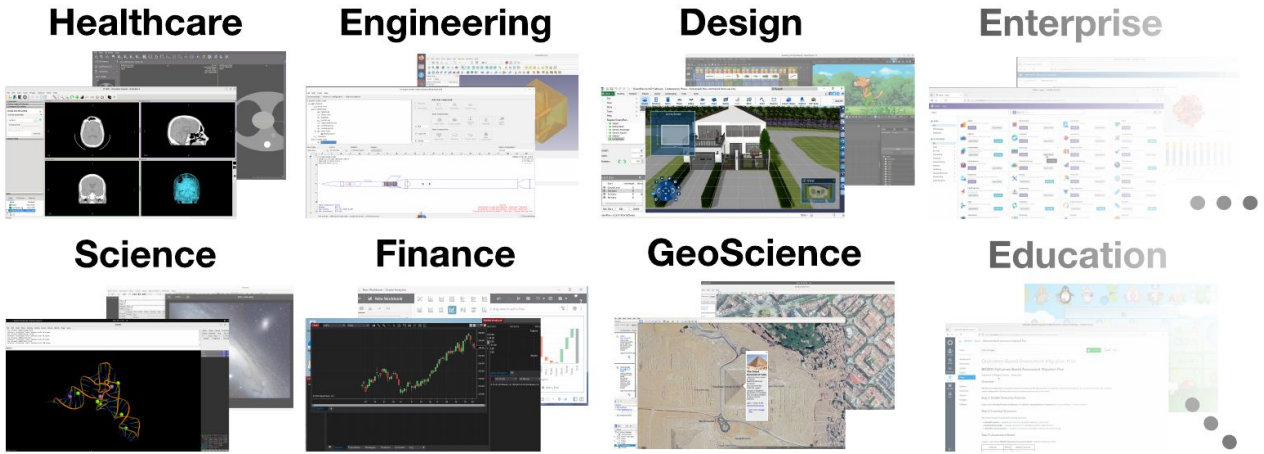
OSWorld 2.0: Modeling a part in SolveSpace

TASK: Recreate a mechanical part from the reference video's dimensions and steps, then save it as a SolveSpace file.



Koh, Sep 2026 · What's the Point of Computer Use Agents? · GPT-6 Astra demo (30×)

CUA-World-Long: One hard task per software



BENCHMARK STATS

200
long tasks

>500
steps often needed

8
quality criteria

EVALUATION CRITERION

Checklist VLM verifier

One long task per application exposes failures across real professional software.

Source: Aggarwal et al., 2026, Fig. 1 (cropped)

Modeling

A VLM predicts actions from visual history

CONTEXT AT STEP 0

“Buy a blue mug < $20”

GOAL

SCREENSHOT 0

A VLM predicts actions from visual history

CONTEXT AT STEP 0

"Buy a blue mug < $20"

GOAL

SCREENSHOT 0

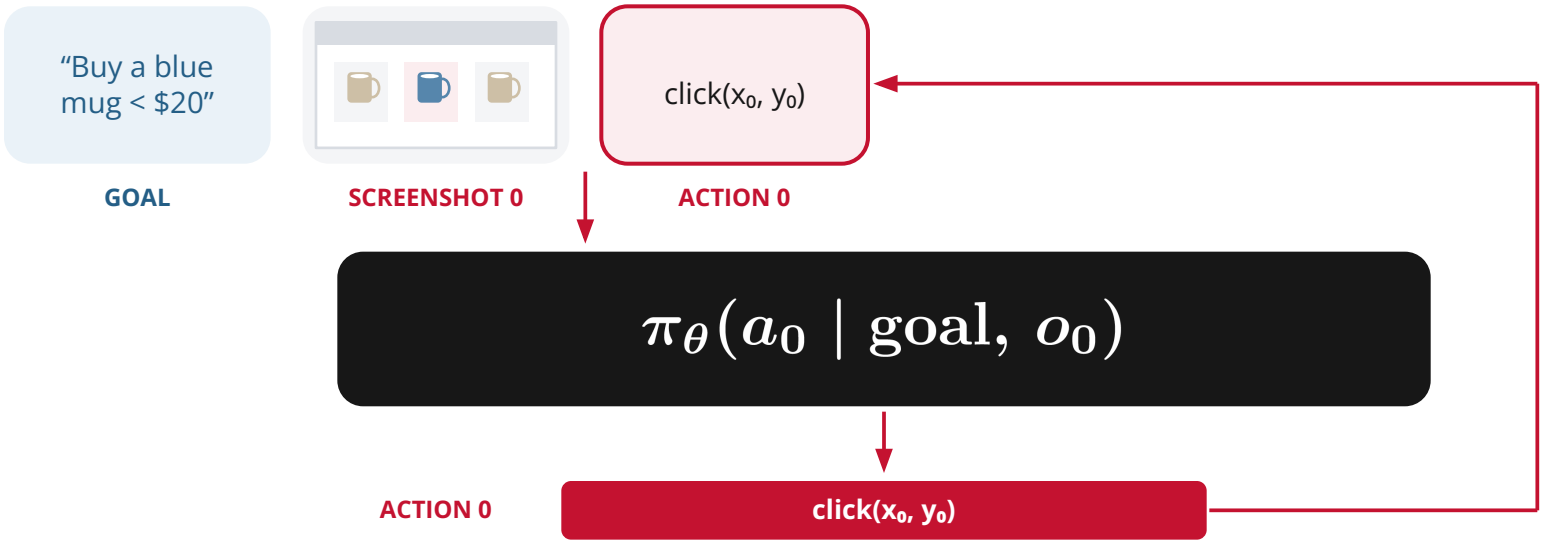
$$\pi_\theta(a_0 \mid \text{goal}, o_0)$$

ACTION 0

click(x_0, y_0)

A VLM predicts actions from visual history

APPEND ACTION 0 TO THE TRAJECTORY



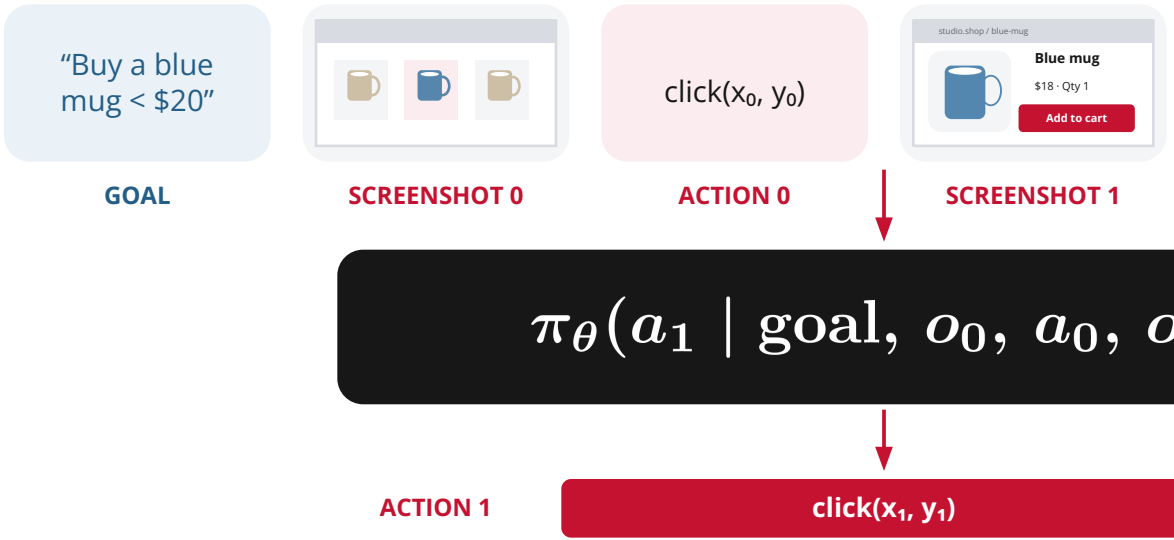
A VLM predicts actions from visual history

CONTEXT AT STEP 1

$$\pi_\theta(a_1 \mid \text{goal}, o_0, a_0, o_1)$$

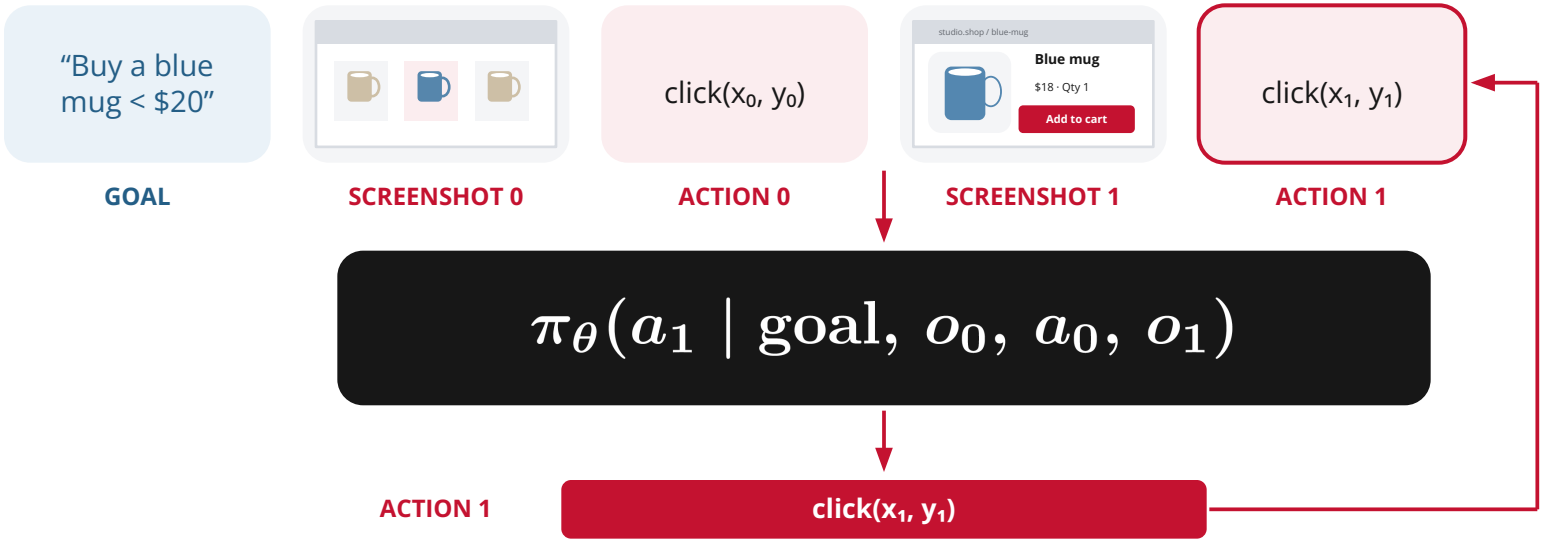
A VLM predicts actions from visual history

CONTEXT AT STEP 1



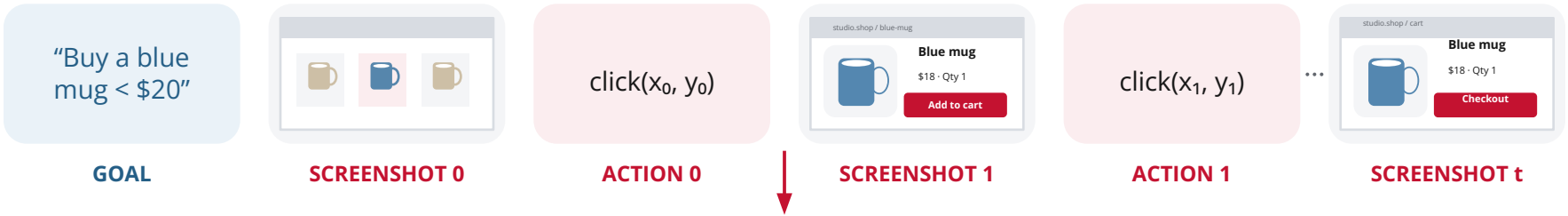
A VLM predicts actions from visual history

APPEND ACTION 1 TO THE TRAJECTORY



A VLM predicts actions from visual history

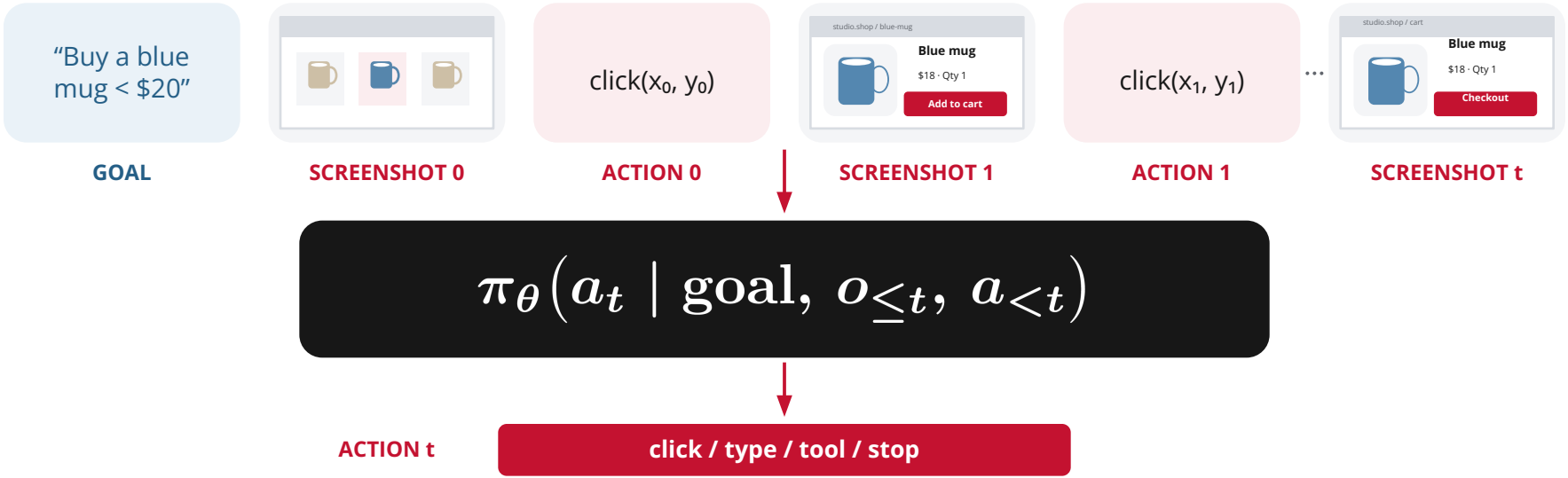
CONTEXT AT STEP t



$$\pi_\theta(a_t \mid \text{goal}, o_{\leq t}, a_{<t})$$

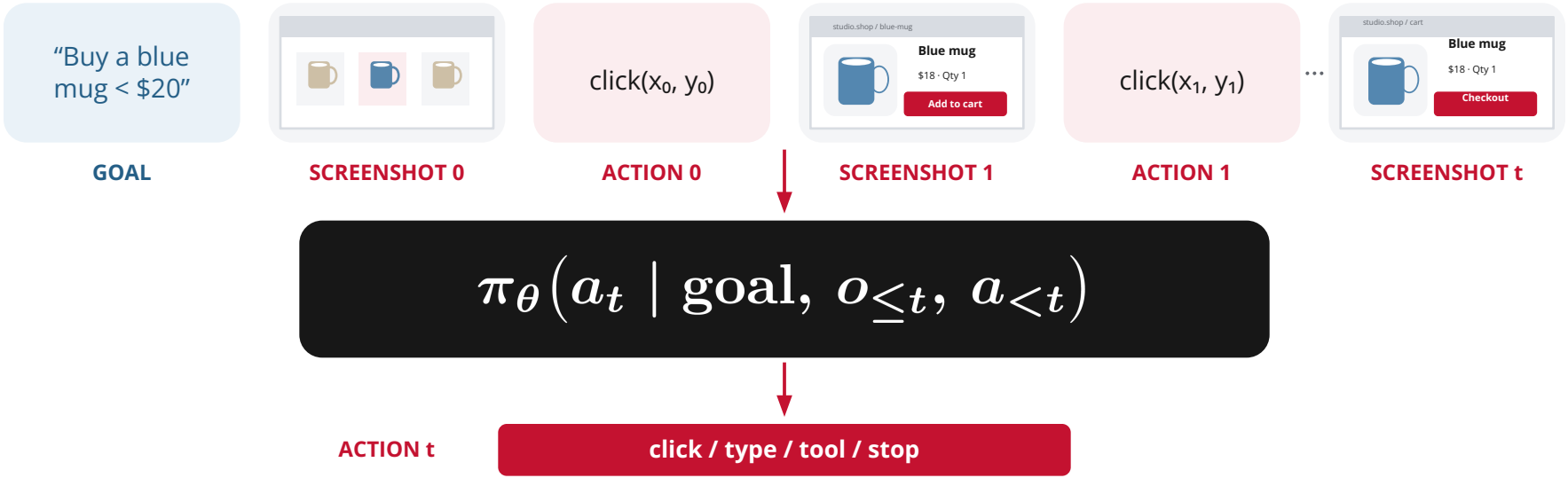
A VLM predicts actions from visual history

CONTEXT AT STEP t



A VLM predicts actions from visual history

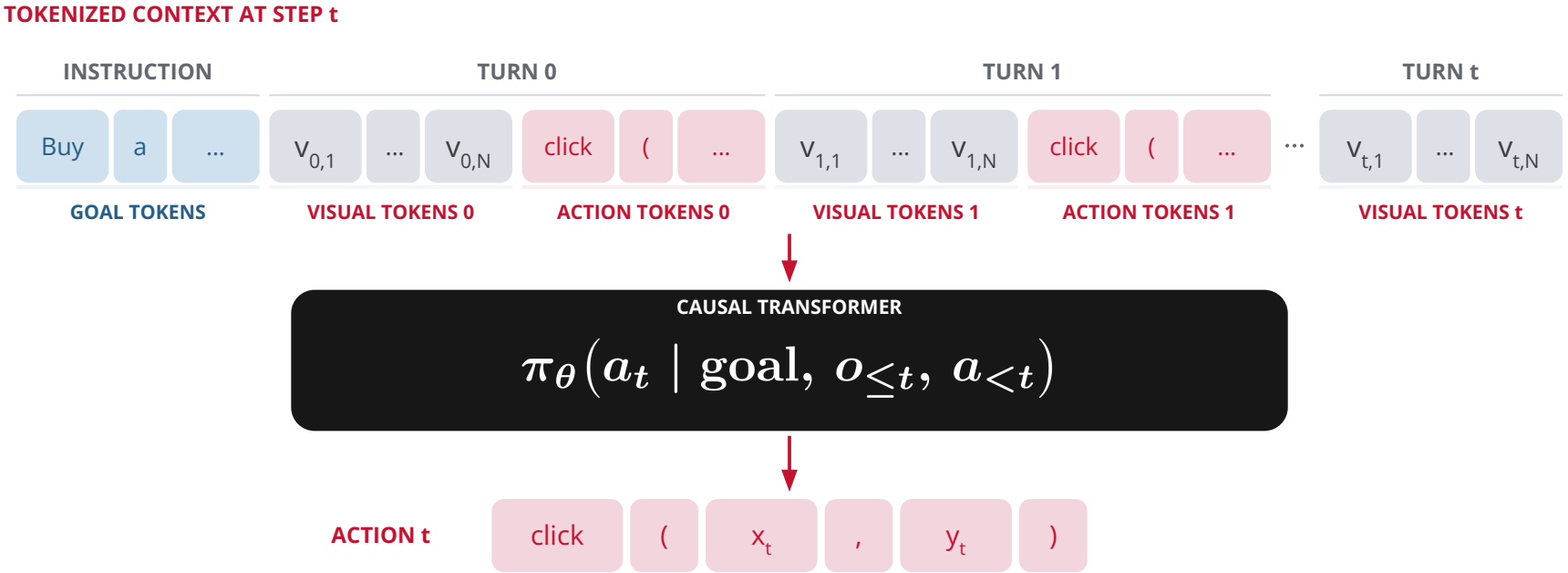
CONTEXT AT STEP t



Interleaved visual and action history lets the policy condition on past progress.

Generic VLM policy · o denotes observations, a actions · tokenization and history compression vary by system

A VLM predicts actions from interleaved tokens



Text → tokenizer + embeddings; screenshots → vision encoder + projector.

Illustrative token splits; visual tokens are continuous embeddings, not vocabulary IDs. Token counts / encoders vary.

Frontier CUAs: Similar observations, different actions

MODEL	ACTION INTERFACE	OUTPUT FORMAT / SCHEMATIC EXAMPLE
GPT-6 Astra	Python / PyAutoGUI or native computer tool	`pyautogui.click(x, y)` `computer_call: actions=[…]`
Fable / Opus 5	Native computer tool calls (one or more per response)	`tool_use: left_click` `input: {coordinate: [x, y]}`
Gemini 3.8 Flash	Computer-use function calls (normalized coordinates)	`functionCall: click` `args: {x, y}  ∈ 0–999`
Qwen 3.8	Function calls (app-defined GUI schema)	`<tool_call><function=…>…` `XML-style tool serialization`[1]
Muse Spark	Scripts + direct GUI actions including action batches	Tool / script output[2] Exact GUI schema is app-specific
Kimi K3	Function calls (app-defined GUI schema)	`tool_calls[].function` `name + JSON arguments`

Output format and schema differs between models (reproducibility is a nightmare).

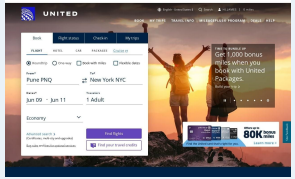
Sources: OpenAI · Anthropic · Google · Qwen[1] · Meta[2] · Moonshot | [1]27B template; [2]1.1 report | Sep 2026

Training

Training Stages

Pre-training

UI ELEMENTS



boxes · labels · OCR
→ grounding

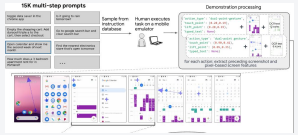
ACTIONS

- click(x, y)
- type(text)
- swipe(...)

click · type · swipe
→ control

Post-training (SFT)

HUMAN DEMOS



goal + human trace
→ behavior cloning

SYNTHETIC DATA

Teacher

o0	a0
o1	a1

agent-generated
traces
→ behavior cloning

Post-training (RL)

RL ROLLOUTS

o → a → ...

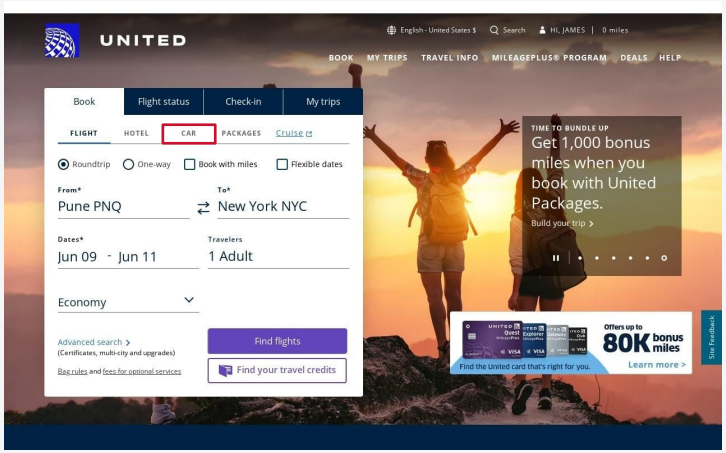
reward
+1

task + verifier
→ task reward

Visuals: Mind2Web (Deng et al., 2023) · AITW (Rawles et al., 2023) · Editable teaching schematics

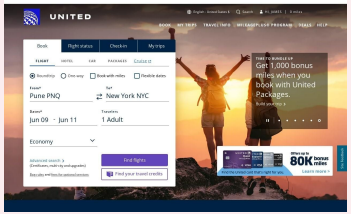
Pre-training: UI elements and actions

UI ELEMENTS / GROUNDING



“CAR” ↔ element box / screen coordinates

ACTION PREDICTION



→ **CLICK**

element: bookCarTab
value: none

Other labels: type(text), swipe(start, end)

Learn what is on the screen, and how to address it.

Deng et al., 2023 · Mind2Web training example; Wu et al., 2025 · OS-Atlas

Post-training (SFT): Imitate useful trajectories

HUMAN DATA

MolmoWebMix (example above)
36K human + 105K synthetic task trajectories

AgentNet
22,625 released task trajectories · 3 OS

SYNTHETIC DATA

AxTree teacher → **Roll out** → **Filter**

Record format (Fig. 2) · 105K synthetic tasks

Retain a useful trajectory → SFT

Same next-action loss; demonstrations can come from humans or agents.

Gupta et al., 2026 · MolmoWeb Fig. 2 excerpt; Table 2; Wang et al., 2025 · OpenCUA Table 2; §2.3

Demonstration data: More scale, richer trajectories

Human and synthetic traces now cover longer, more diverse workflows.

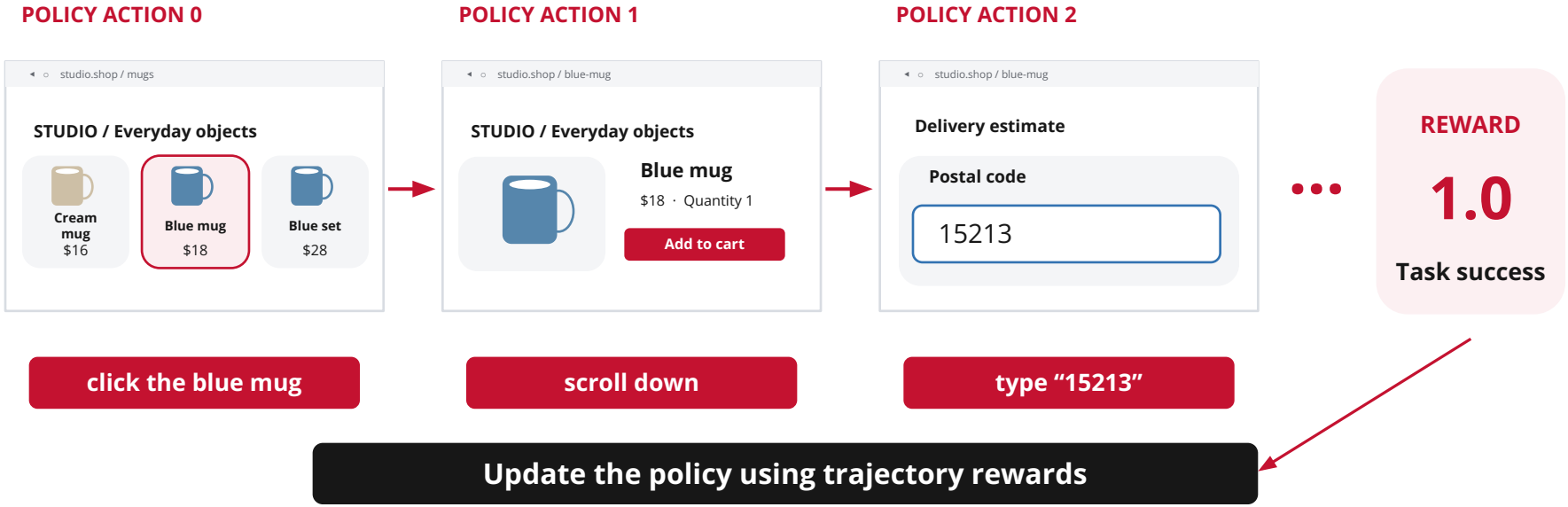
2023	2023	2024	2024	2024	2025	2026
Mind2Web	**AITW**	**WebLINX**	**Android Control**	**GUI Odyssey**	**AgentNet**	**MolmoWeb Mix**
2,350 tasks 137 websites	715K episodes 30K instructions	100K interactions 2.3K demos	15,283 demos 833 apps	7,735 episodes cross-app	22,625 tasks 18.6 steps per trajectory	36K human 105K synth. task trajectories
WEB	MOBILE	WEB	MOBILE	MOBILE	DESKTOP	WEB

Scale has increased over time; data type also shifts towards focusing on longer, cross-app, multi-turn supervision.

Mind2Web · AITW · WebLINX · AndroidControl · GUI Odyssey · AgentNet · MolmoWeb

Post-training (RL): Learn from scored rollouts

Goal: Find a blue mug, check delivery, and add it to the cart.

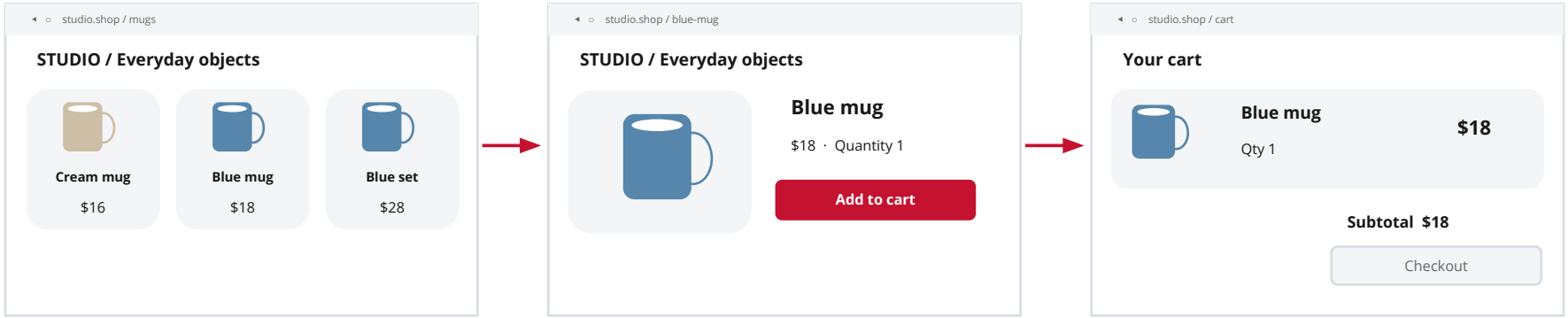


Generate experience → score the outcome → improve the policy.

Illustrative rollout · CUA-Gym / Gym-Anything: Interactive tasks and verifiers

RL Training

TASK: Add one blue mug under $20 to the cart; do not check out.



history + screen

ACTION 0 click product

history + screen

ACTION 1 add to cart

history + screen

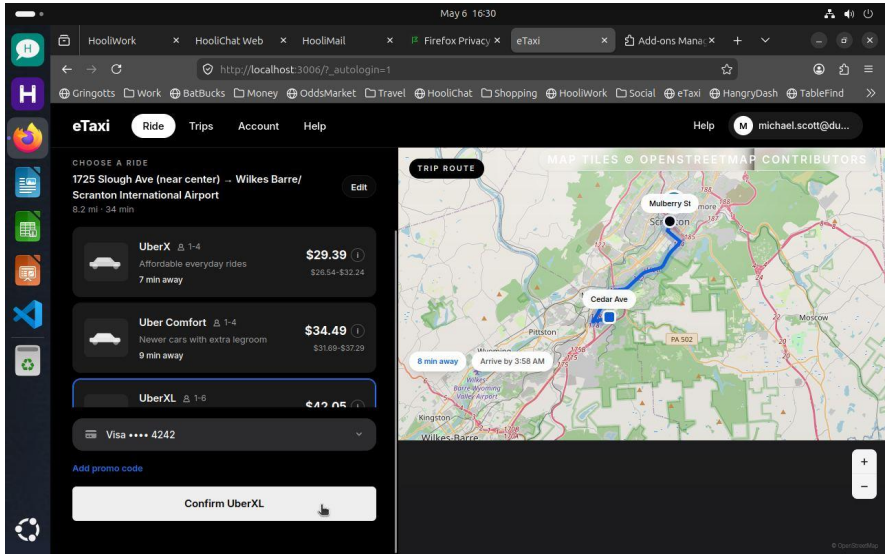
ACTION 2 verify / stop

SFT learns next action prediction; RL evaluates trajectories based on outcomes.

Teaching example · screenshots, actions, and optional rationale labels are interleaved

We need simulated environments for RL

TASK: Book an airport ride for my upcoming Jamaica flight.



REAL MONEY

Repeated attempts can create duplicate paid bookings.

REAL PEOPLE

A booking can dispatch a driver; exploration affects other people.

NO CLEAN RESET

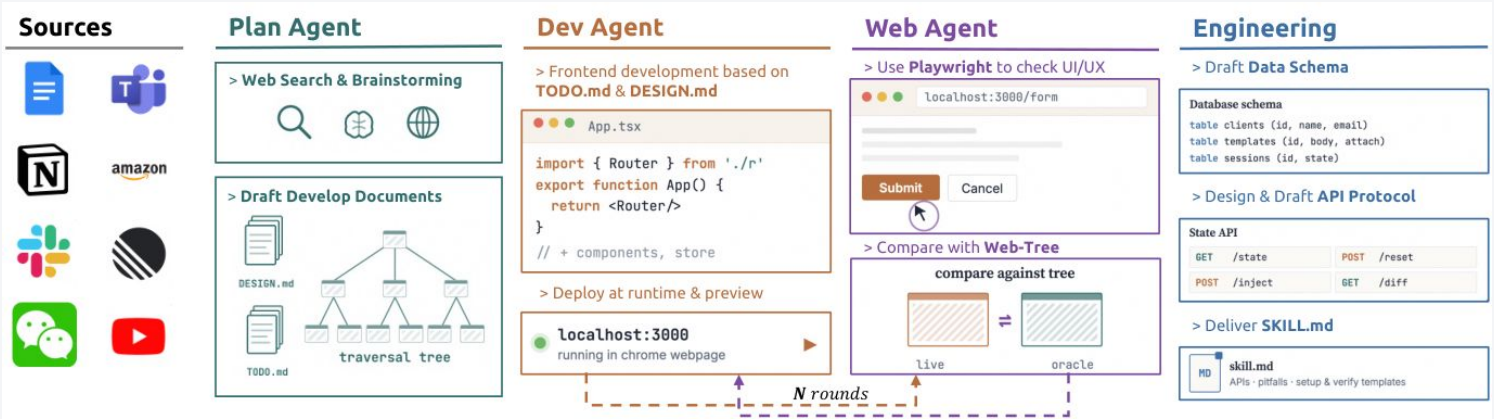
Cancellation may incur a fee; spent time cannot be restored.

Explore in resettable replicas, not repeated real world interactions.

MyPCBench · Jang et al., 2026 · Task bounded_action-f041 · Published trajectory, step 53

CUA-Gym: Build reusable, resettable mock applications

94 MOCK WEB APPS / PLAN → IMPLEMENT → TEST



Inject initial state

Inspect final state

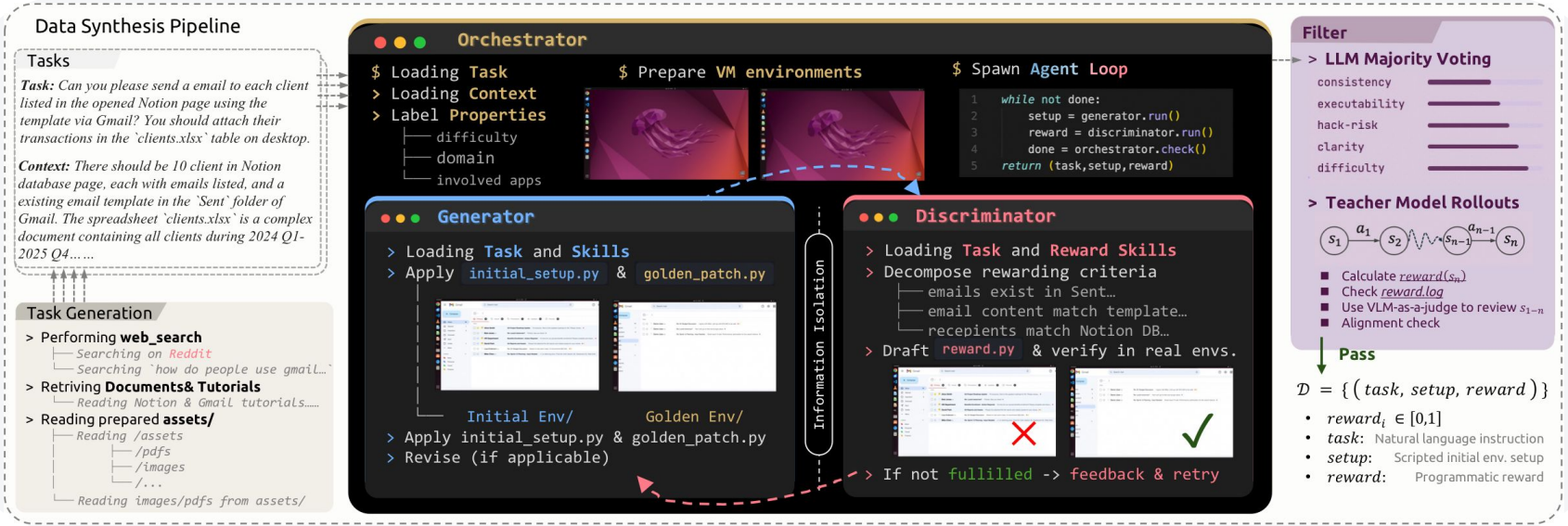
Reset each session

One mock app supports many tasks through controlled, isolated state.

Wang et al., 2026 · CUA-Gym Fig. 2; §2.2

CUA-Gym: Co-generate tasks, states, and rewards

~32K VERIFIED TUPLES / 110 ENVIRONMENTS

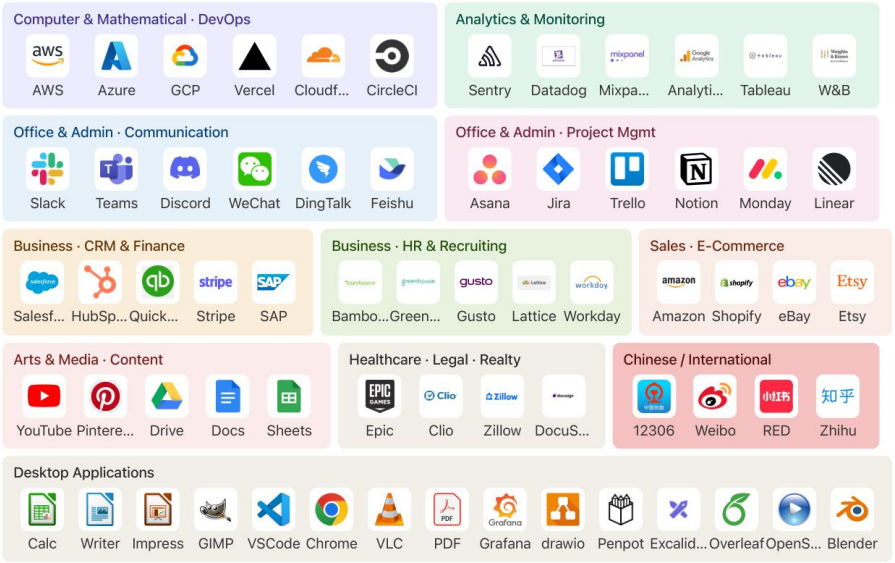


Generate the world and its reward independently—then test that they agree.

Wang et al., 2026 · CUA-Gym Fig. 1; §2.1

CUA-Gym: How much training data?

94 MOCK WEB APPS + 16 DESKTOP APPS



110

environments

32,112

verified RLVR tuples

38%

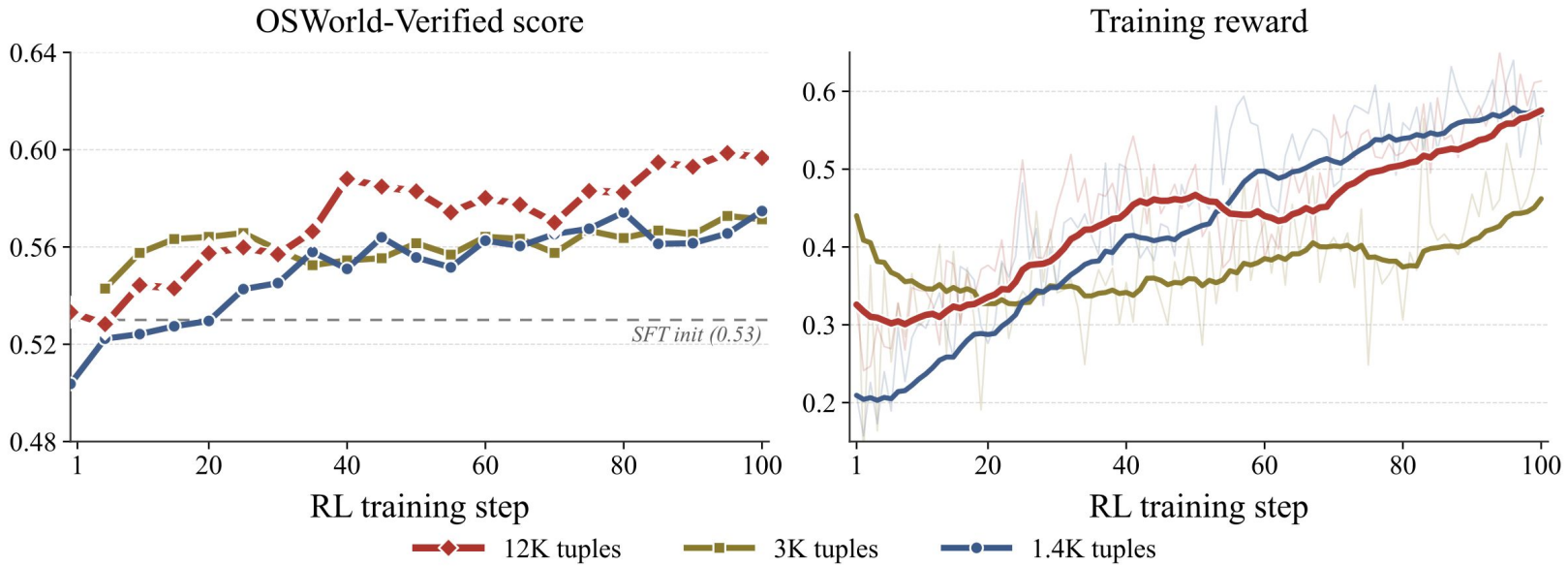
cross-application tasks

A broad environment pool supports many independently verified tasks.

Wang et al., 2026 · CUA-Gym Fig. 8(a); §2.3

CUA-Gym: Verified rollouts improve the policy

3,578 SFT DEMOS → GSPO ON 10,858 VERIFIED TUPLES

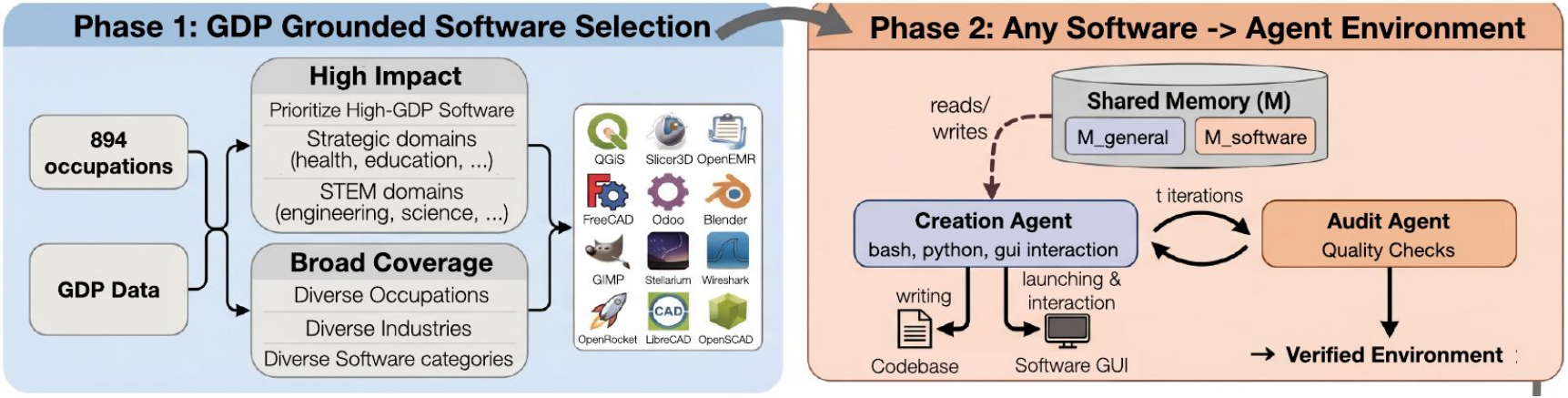


OSWorld Performance: 62.2 → 72.6%

Wang et al., 2026 · Fig. 9 data-scale ablation; §3 main results
OSWorld-Verified · GUI calls · 100 model steps · programmatic checks

Gym-Anything: Create and audit real software setups

SELECT SOFTWARE / CREATE ↔ AUDIT



200
software applications

3
operating systems

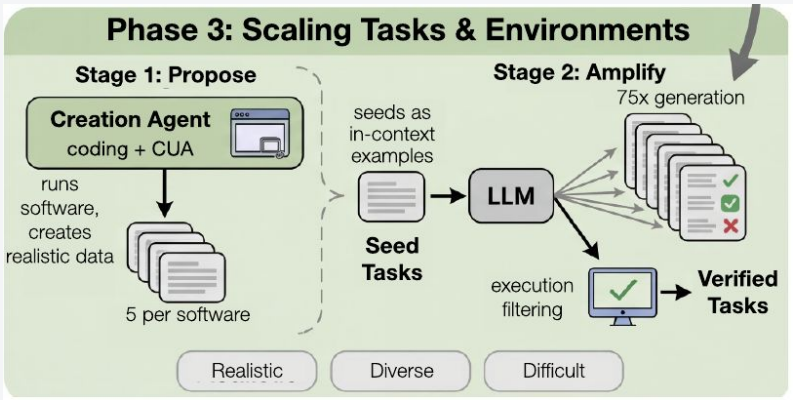
10K+
tasks in CUA-World

The creation agent builds, an independent audit verifies the evidence.

Aggarwal et al., 2026 · Gym-Anything Fig. 2, phases 1–2; §3

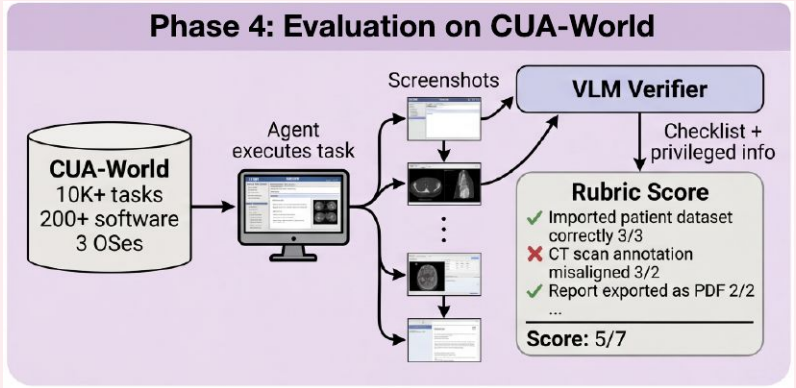
Gym-Anything: Scale tasks, then verify trajectories

PROPOSE + AMPLIFY



Execute seed tasks
Amplify with an LLM → filter

CHECKLIST-BASED VERIFICATION

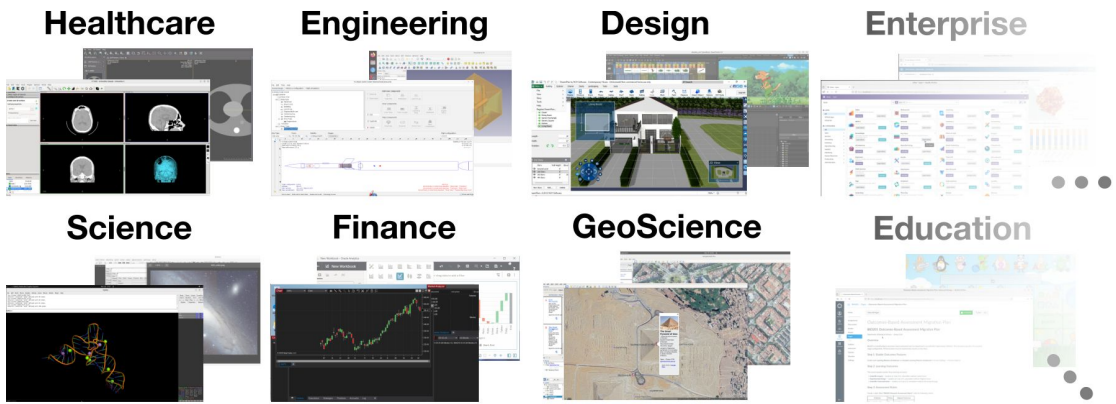


Privileged evidence + weighted checklist
Partial credit + integrity checks

Aggarwal et al., 2026 · Gym-Anything Fig. 2, phases 3–4; §4.1

Gym-Anything: Environments across real software

REAL INSTALLED SOFTWARE / LINUX · WINDOWS · ANDROID



200

software applications

12,103

tasks and environments

22 / 22

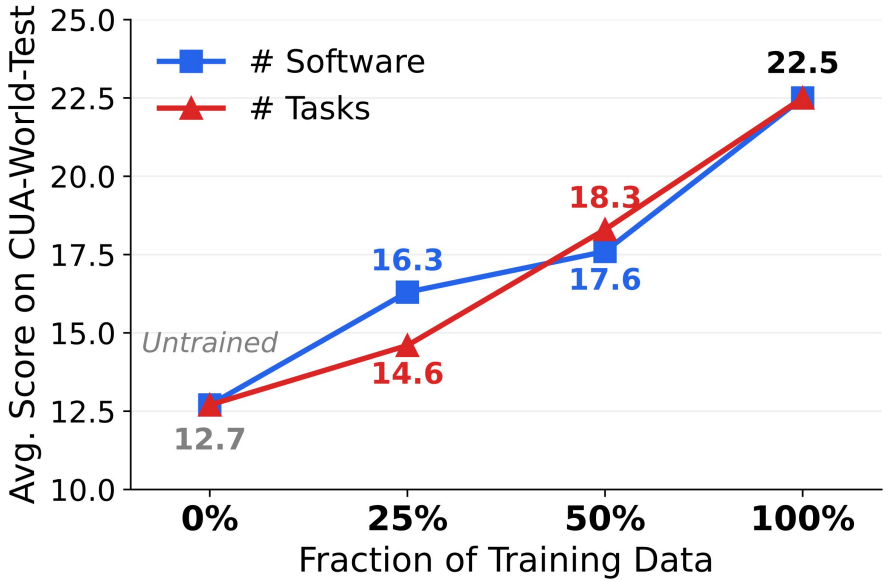
SOC occupation groups covered

The generator creates task-ready setups across professional software.

Aggarwal et al., 2026 · Gym-Anything Fig. 1; §5

Gym-Anything: Distill successful trajectories

KIMI-K2.5 TEACHER → QWEN3-VL-2B-THINKING STUDENT



~2,000

successful SFT trajectories

12.7 → 22.5

average checklist score / 100

1.6 → 4.4%

perfect-checklist pass rate

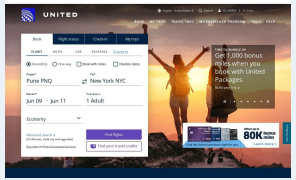
More software and more tasks both help supervised post-training.

Aggarwal et al., 2026 · Fig. 6(a); Table 3; §6.1–7.1
CUA-World-Test · GUI actions · 200 steps · Gemini 3 Flash checklist judge

Training Stages

Pre-training

UI ELEMENTS



boxes · labels · OCR
→ grounding

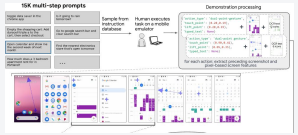
ACTIONS

- click(x, y)
- type(text)
- swipe(...)

click · type · swipe
→ control

Post-training (SFT)

HUMAN DEMOS



goal + human trace
→ behavior cloning

SYNTHETIC DATA

Teacher

o0	a0
o1	a1

agent-generated
traces
→ behavior cloning

Post-training (RL)

RL ROLLOUTS

o → a → ...

reward
+1

task + verifier
→ task reward

Strong CUA

Visuals: Mind2Web (Deng et al., 2023) · AITW (Rawles et al., 2023) · Editable teaching schematics

Summary

1. **CUAs operate directly on the GUI**
   They observe and act within the same interface as humans.

2. **Evaluations should target the right metrics and level of behavior**
   Static actions, end-to-end tasks, and long horizon workflows are measured differently.

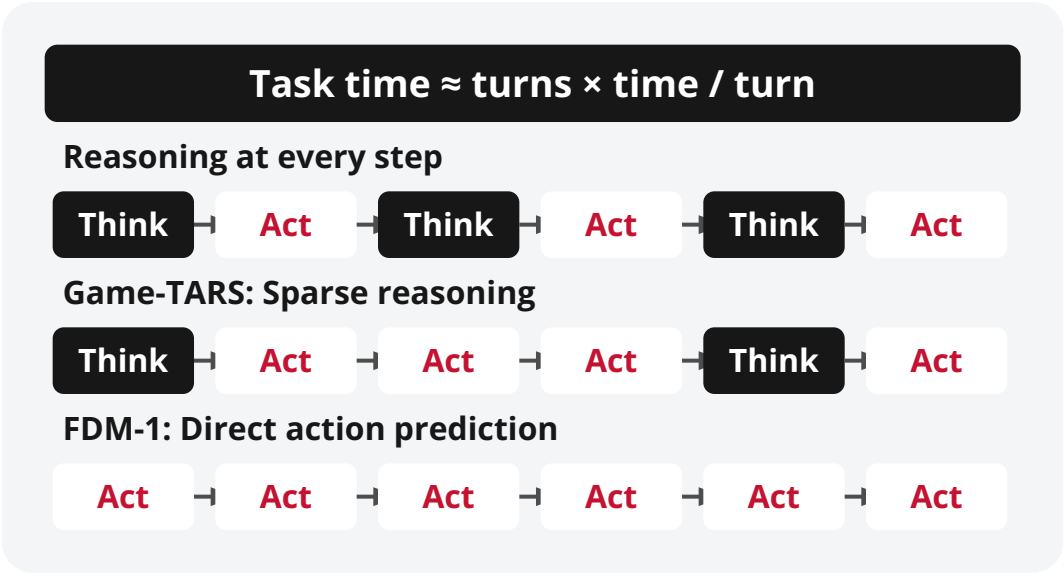
3. **Models predict actions from interleaved visual history**
   CUAs are vision-language models that emit GUI actions, some also use code and tools.

4. **Pretraining → SFT → RL**
   Models are trained through large scale pretraining, SFT on human annotated and synthetic trajectories, and RL within simulated environments.

What’s remaining?

Speed + Cost

FASTER CUAS REQUIRE BOTH SYSTEMS AND MODEL IMPROVEMENTS



Long contexts are expensive

Screenshots and reasoning add prefill and generation cost.

Game-TARS: Sparse Thinking

Reason at key decisions; act directly on routine steps.

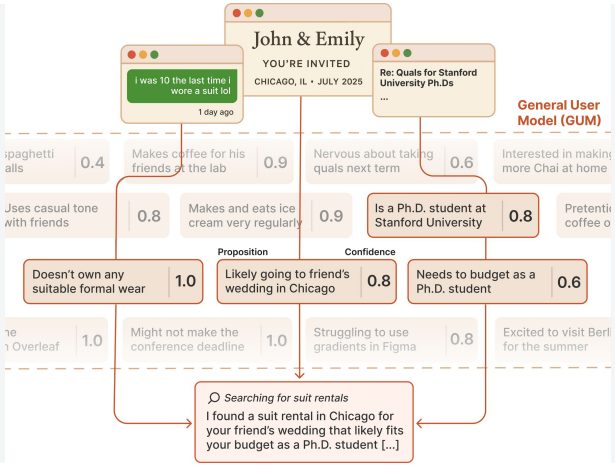
Fewer / cheaper turns

Improve planning and batching; explore direct action models (FDM-1).

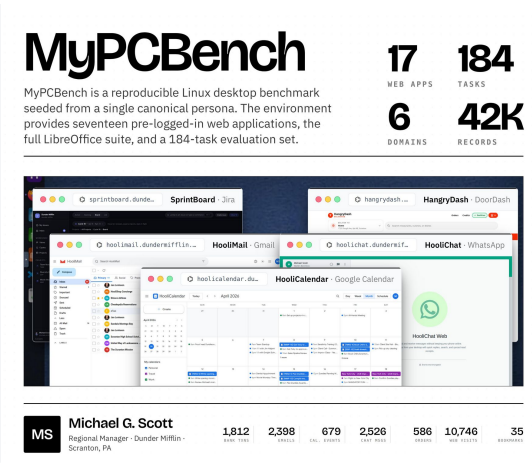
Game-TARS · ByteDance Seed, 2025, §3.3 / §4.3.2 · FDM-1 · Standard Intelligence, 2026

Personalization: Adapt to the user

GUM: INFER THE USER'S CONTEXT



MYPCBENCH: A PERSONAL DIGITAL LIFE



Personal context changes

Use the user's files, history and preferences—not generic defaults.

Memory must be revisable

GUM retrieves and updates confidence-weighted user facts.

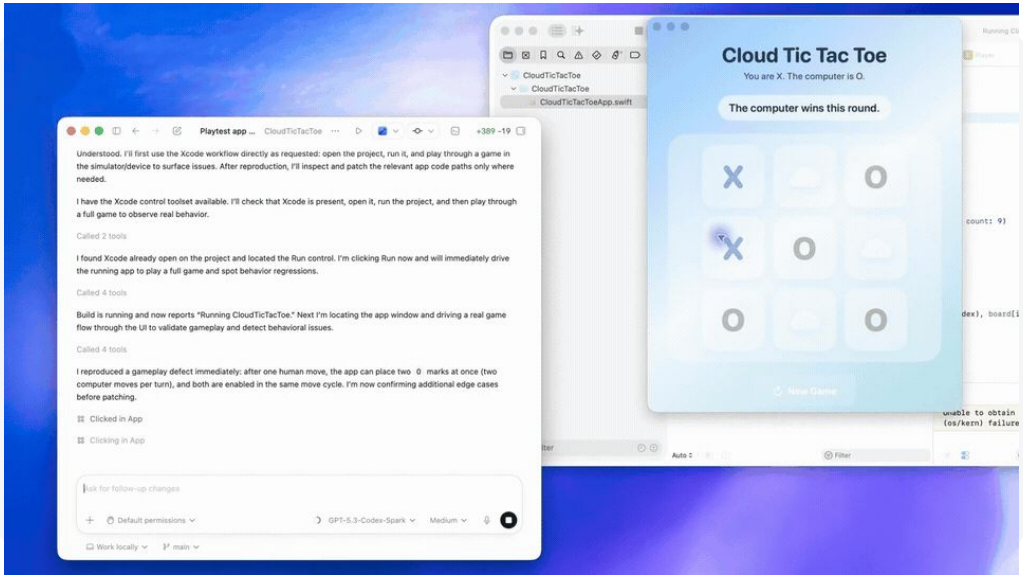
User control is essential

MyPCBench tests personal context. Consent and forgetting are essential.

GUM · Shaikh et al., 2025, Fig. 1 · MyPCBench · Jang et al., 2026, project overview

Infrastructure + UX

CODEX: BACKGROUND COMPUTER USE



Don't block the user

Run in isolated background sessions while their work continues.

Collaborate in real time

Show progress; accept corrections, approval and interruption.

Novel UI / UX for CUAs

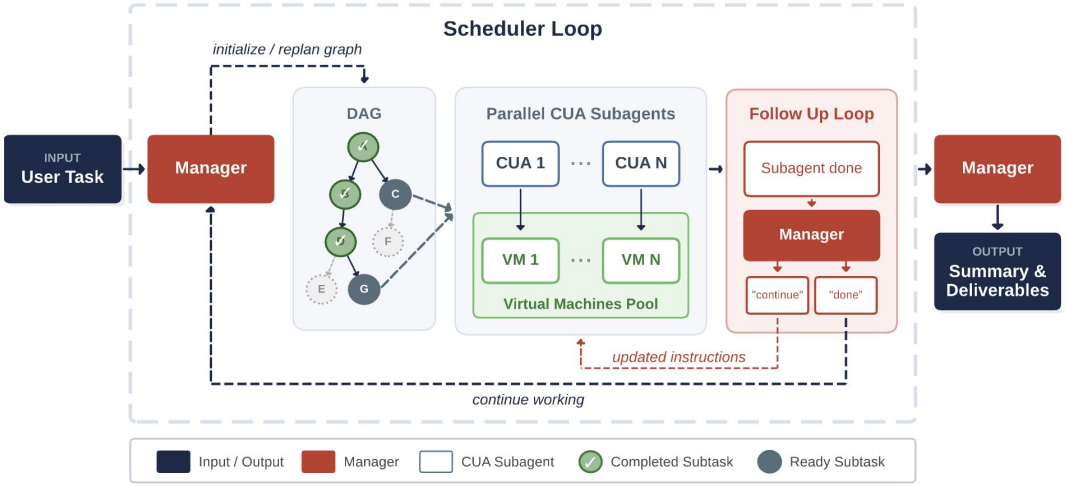
Much to be explored still!

Proactivity needs consent, visibility, and an easy off switch.

OpenAI, 2026 · Codex for (almost) everything · demo excerpt

Multi-Agent Systems

MACU: PLAN → PARALLELIZE → REPLAN



Divide work by dependencies

MACU dispatches ready subtasks to CUAs in isolated environments.

Share evidence, then replan

Hand off files and results; verify work and recover from failures.

Beyond homogeneous CUAs

Combine CUA, code and tool agents; control conflicts and overhead.

Parallelism helps only when coordination preserves correctness.

MACU · Koh, Salakhutdinov & Fried, 2026, Fig. 1 / §3 (adapted)

Thanks!

Appendix: Benchmarks

WEB + STATIC EVALUATION

MiniWoB / MiniWoB++
WebShop
ScreenSpot-Pro
Mind2Web
Android in the Wild
AndroidControl
WebArena
VisualWebArena
WebVoyager
Online-Mind2Web
WorkArena / WorkArena++

DESKTOP, MOBILE + LONG HORIZON

OSWorld / Verified
WindowsAgentArena
AndroidWorld
MobileWorld
Odysseys
OSWorld 2.0
MyPCBench
WeaveBench

AgentNetBench and CUA-World-Long share their papers with training datasets/environments (p. 99).

Appendix: Modeling / Training

MODELS + GROUNDING

OS-Atlas

UI-TARS

Qwen-UI-Agent

Qwen3.6-27B (model card)

Game-TARS

DATA + TRAINING ENVIRONMENTS

OpenCUA / AgentNet / AgentNetBench

MolmoWeb / MolmoWebMix

CUA-Gym

Gym-Anything / CUA-World

Grounding and demonstration datasets that also serve as static benchmarks are listed on p. 98.

Appendix: Systems

AGENT LOOPS + ORCHESTRATION

ReAct

Interleaved reasoning and actions

MACU: Multi-Agent Computer Use

Orchestrating multiple computer-use agents

GUM: General User Models

User modeling from computer activity

FRAMEWORKS + EFFICIENCY

BrowserGym ecosystem

Shared environments and experiment tooling

Cua (open-source framework)

Computer-use environments and infrastructure

FDM-1 (technical blog)

High-frequency computer action modeling

Papers and technical resources for agent orchestration, personalization, infrastructure, and efficiency.

Appendix: Products

OPENAI + ANTHROPIC

Codex Computer Use (OpenAI blog)

Operator (OpenAI launch)

OpenAI computer-use API guide

Claude computer use (launch blog)

Developing computer use (Claude blog)

Claude computer-use tool docs

OTHER PRODUCTS + RELEASES

Grok Bot (xAI launch)

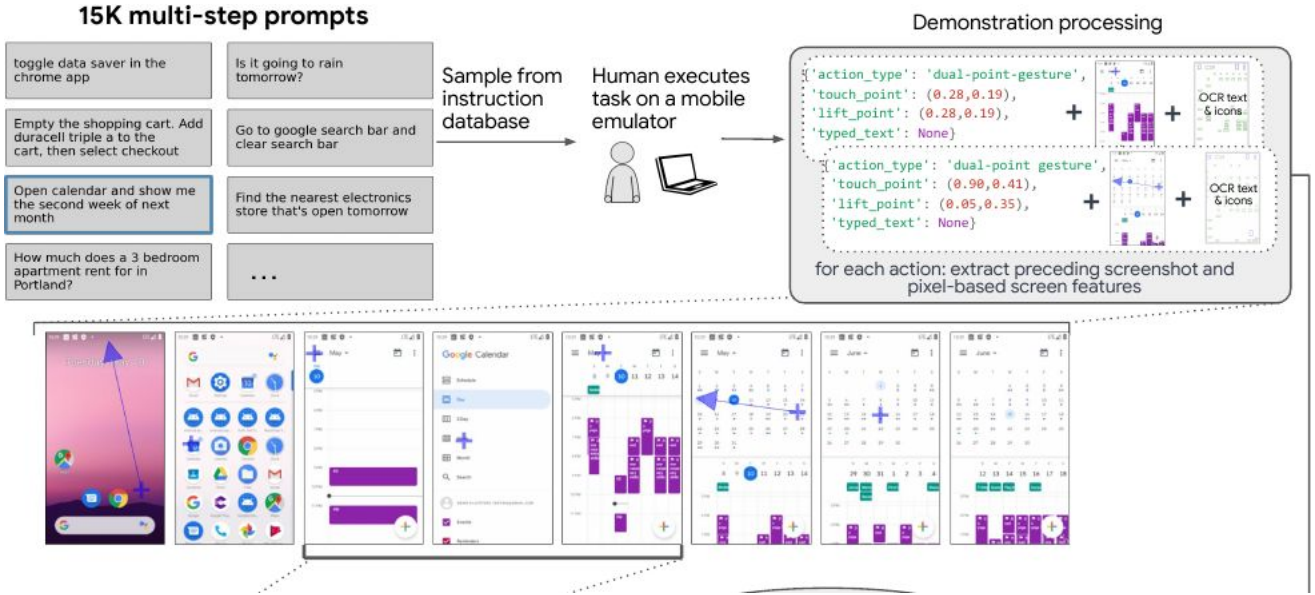
Gemini computer-use API guide

Muse Spark 1.1 (Meta release)

Navigator n2 (Yutori release)

Official product announcements, technical blogs, and API documentation. Operator is a historical launch.

Android in the Wild: Scale without end-state checks



BENCHMARK STATS

715K
episodes

5.69M
examples

30,378
unique prompts

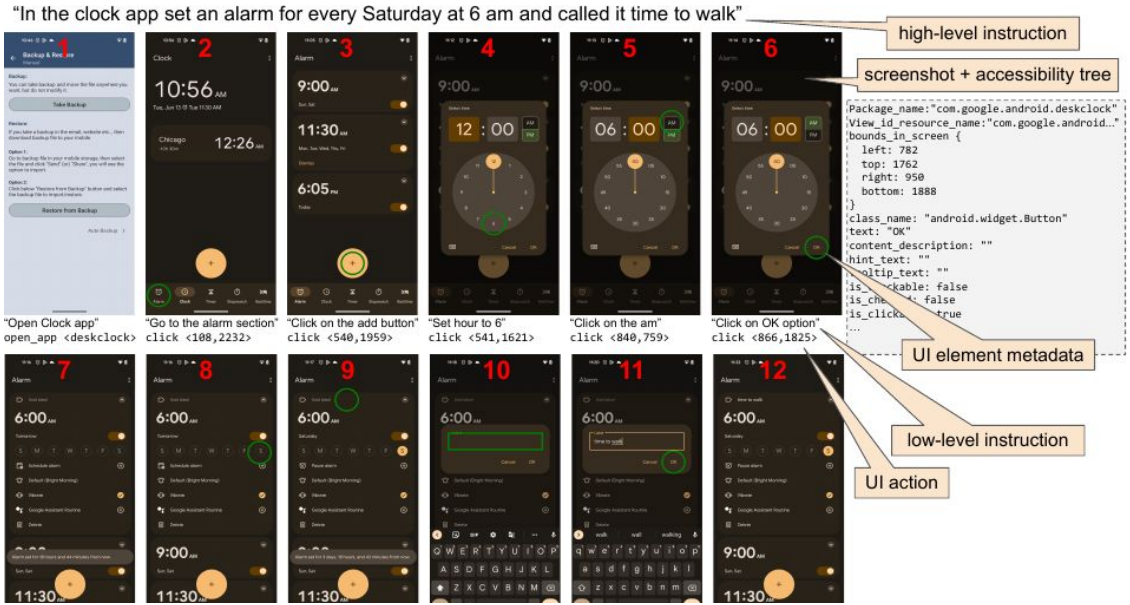
EVALUATION CRITERION

Offline action matching

Scale buys coverage; the evaluator still matches recorded actions.

Source: Rawles et al., 2023, Fig. 1

AndroidControl: Diverse apps, paired instructions



BENCHMARK STATS

15,283
demonstrations

14,548
unique tasks

833
Android apps

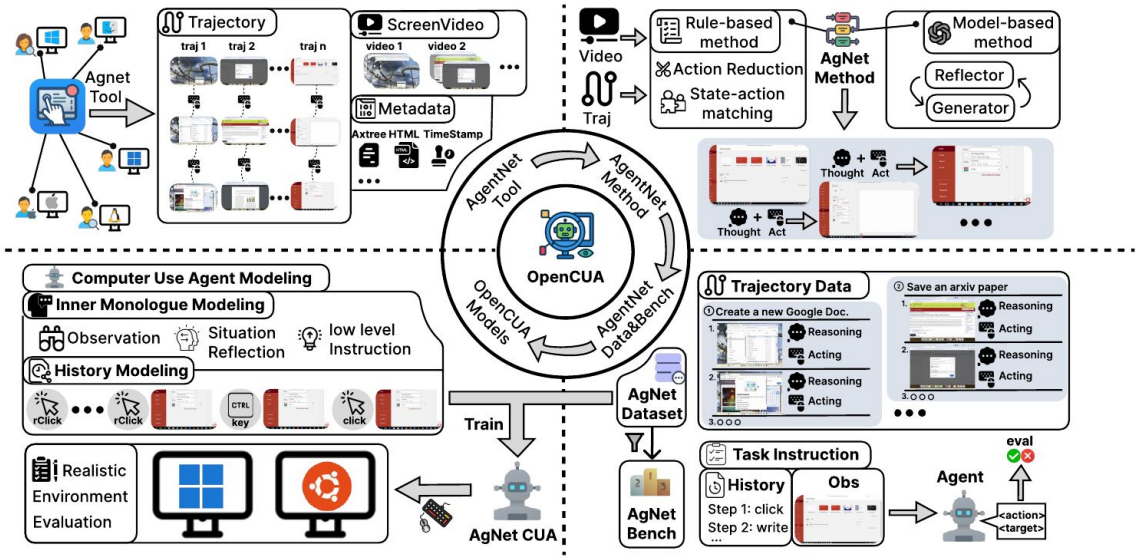
EVALUATION CRITERION

Offline action matching

Paired high- and low-level instructions separate planning from motor control.

Source: Li et al., NeurIPS 2024, Fig. 1

AgentNetBench: An offline desktop proxy



BENCHMARK STATS

100
tasks

17.63
average steps

2,143
total actions

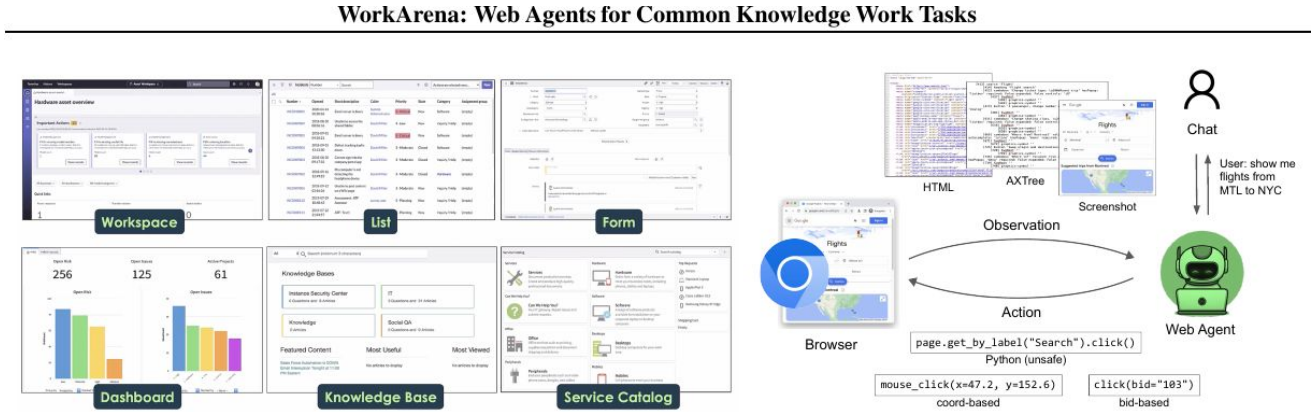
EVALUATION CRITERION

Offline step matching

AgentNetBench turns human desktop trajectories into a fast offline proxy.

Source: Wang et al., NeurIPS 2025, Fig. 2 (cropped)

WorkArena: Enterprise tasks with validators



BENCHMARK STATS

33
task templates

19,912
instances

ServiceNow
environment

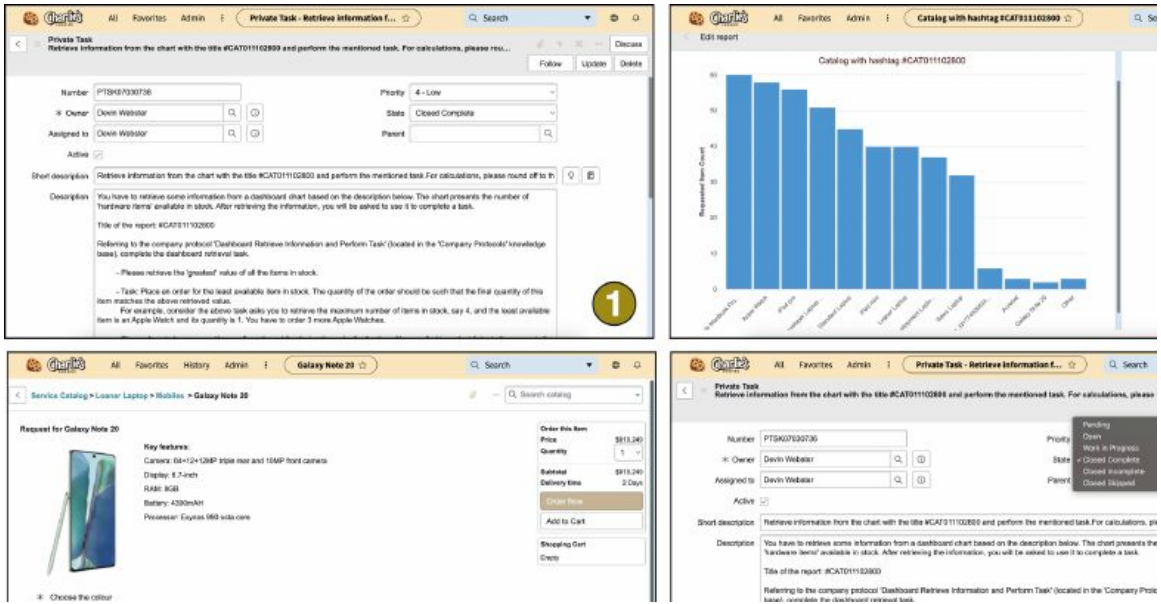
EVALUATION CRITERION

Programmatic checks

Parameterized tasks make enterprise state validation reproducible.

Source: Drouin et al., ICML 2024, Fig. 1

WorkArena++: Compositional enterprise workflows



BENCHMARK STATS

682
composite tasks

5
skill families

10
company themes

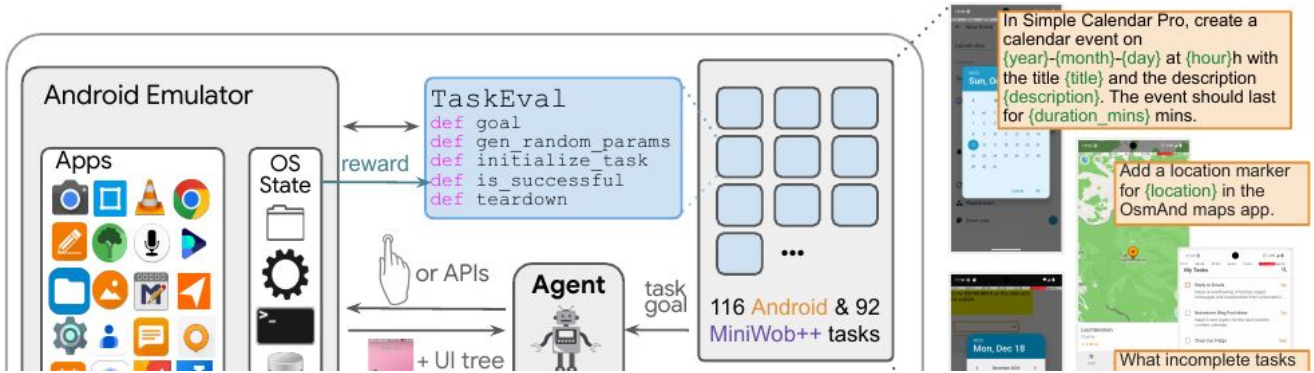
EVALUATION CRITERION

Compositional end state

Composition converts atomic UI skills into knowledge-work workflows.

Source: Boisvert et al., NeurIPS 2024, Fig. 1

AndroidWorld: Parameterized mobile state



BENCHMARK STATS

116
task families

20
applications

∞
parameterized instances

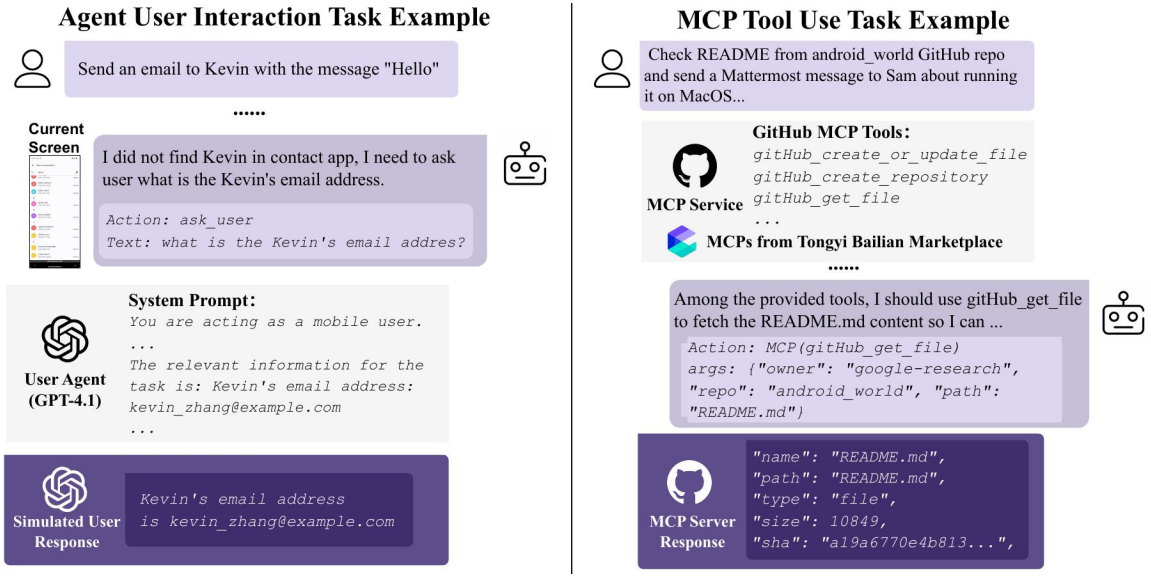
EVALUATION CRITERION

System-state reward

Task generators and system-state rewards create repeatable mobile episodes.

Source: Rawles et al., ICLR 2025, Fig. 1

MobileWorld: Cross-app, user, and MCP tasks



BENCHMARK STATS

201
tasks

27.8
average steps

62.2%
cross-app tasks

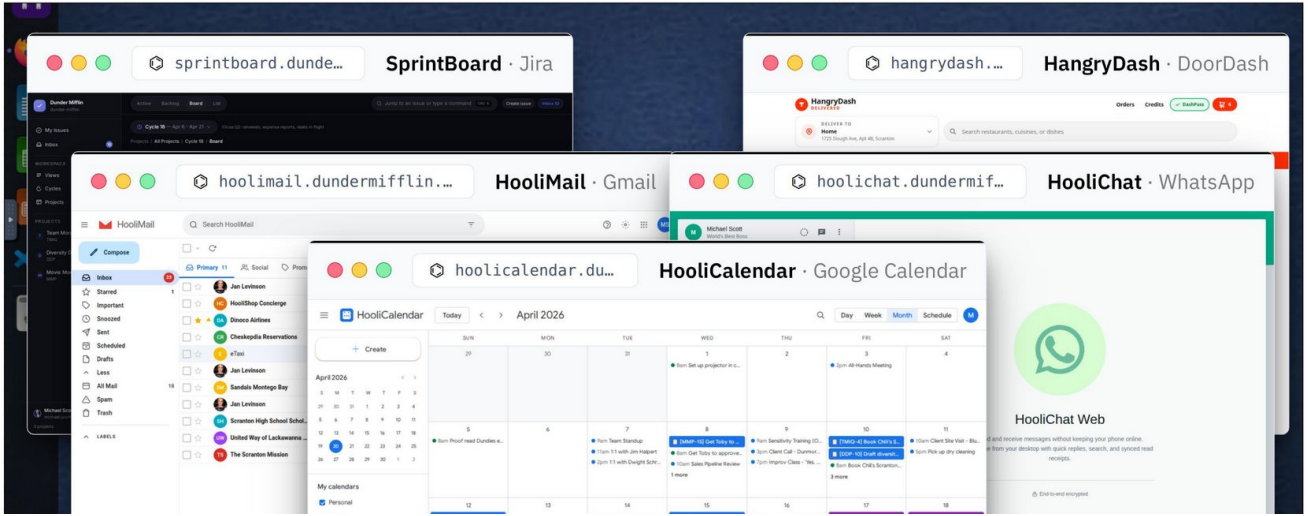
EVALUATION CRITERION

Programmatic + hybrid

Modern mobile use mixes GUI control, clarification, and structured tools.

Source: Kong et al., ACL 2026, Fig. 2 (cropped)

MyPCBench: One person's cross-app digital life



BENCHMARK STATS

184
tasks

17
logged-in apps

68%
multi-app tasks

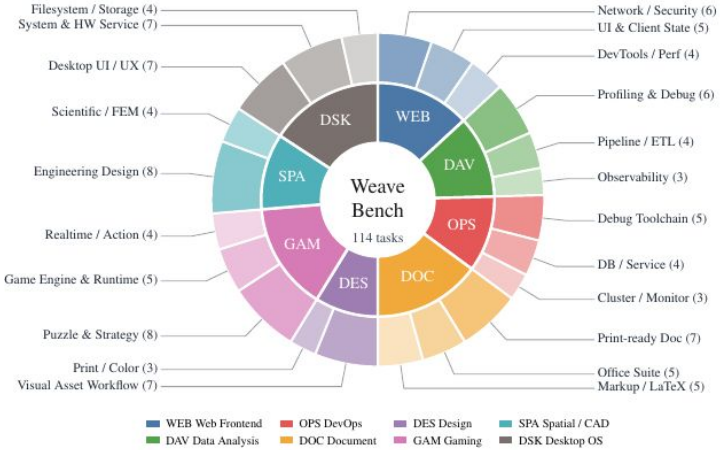
EVALUATION CRITERION

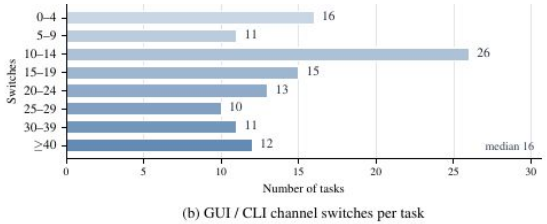
Per-rubric LLM judge

One seeded identity turns separate apps into one reproducible digital life.

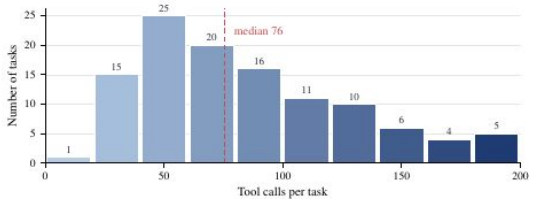
Source: Jang et al., arXiv 2026, Fig. 1 (cropped)

WeaveBench: Hybrid GUI + CLI trajectories





(b) GUI / CLI channel switches per task



BENCHMARK STATS

114
tasks

16
median switches

76
median tool calls

EVALUATION CRITERION

Trajectory-aware judge

Outcome-only grading can overcredit agents that take invalid shortcuts.

Source: Li et al., EMNLP 2026, Fig. 3