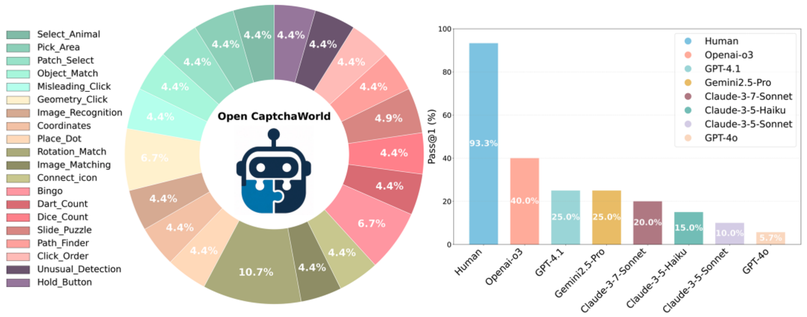
Open CaptchaWorld: A Comprehensive Web-based Platform for Testing and Benchmarking Multimodal LLM Agents

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Yaxin Luo*, Zhaoyi Li*, Jiacheng Liu, Jiacheng Cui, Xiaohan Zhao, Zhiqiang Shen†

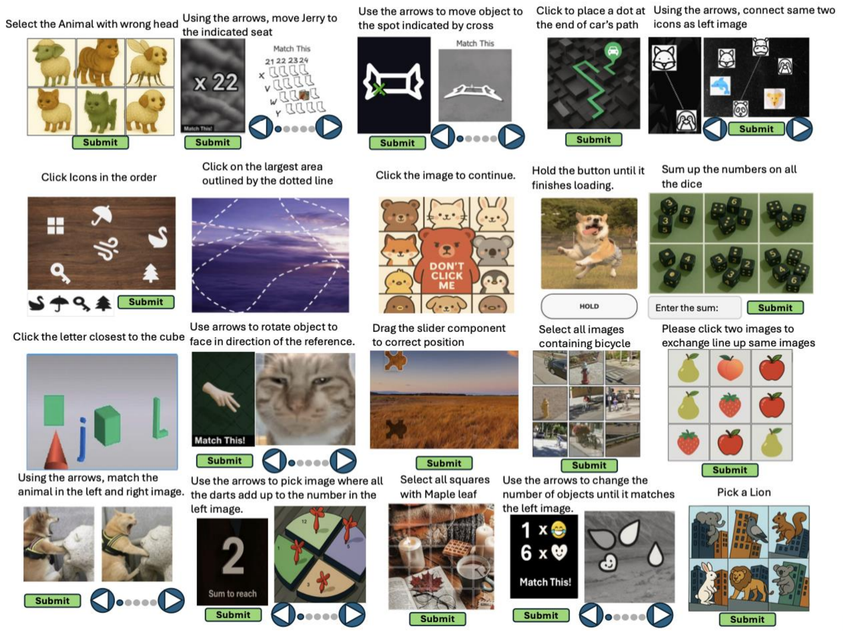
NEURAL INFORMATION PROCESSING SYSTEMS

Open CaptchaWorld



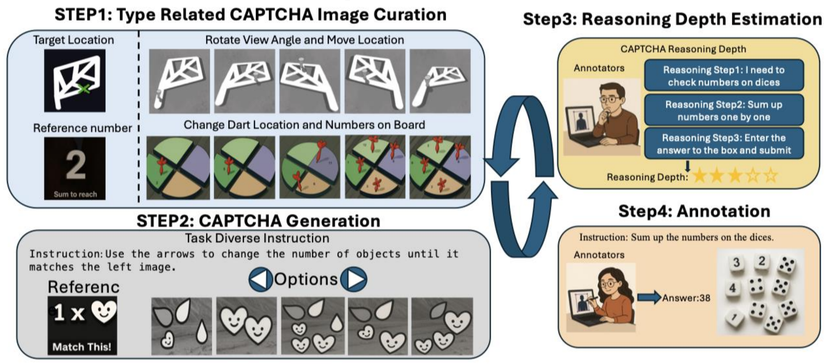
Open CaptchaWorld is a web-based benchmark of 20 modern CAPTCHA types (225 puzzles) to evaluate whether multimodal LLM agents can perceive → reason → act in interactive browser loops. Humans score 93.3%, while the best agent tops out at 40.0%, exposing a big gap.

Some Examples from benchmark



Zero-shot browser-use agents observe screenshots and issue clicks/drag/typing until submit; each run covers all types exactly once.

Dataset Curation Pipeline:



Step 1 — Curation Collect images per CAPTCHA type; vary view, location, and numeric/text cues to make controlled variants.
Step 2 — Generation Write instructions and build web UIs (click/drag/slider/type); add distractors; ensure one unique solution.
Step 3 — Reasoning Depth Annotators decompose the minimal perception→reasoning→action steps; assign a depth score.
Step 4 — Annotation Solve each instance and record the verified answer (plus action trace when needed).

Main Results

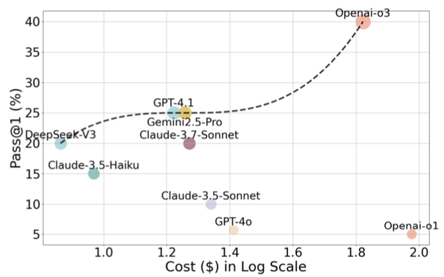
Main Table that compares the different MLLM backbones equipped with Browser-Use agent framework

Solver Type	MLLM Backbone	Pass@1 (%)	Cost ($)
Human	–	93.30	-
Browser Use Agents	GPT-4o	5.7	25.8
	GPT-4.1	25.0	16.7
	Claude-3.7-Sonnet	20.0	18.7
	Gemini2.5-Pro	25.0	18.1
	Openai-o3	40.0	66.4
	Claude-3.5-Haiku	15.0	9.3
	Claude-3.5-Sonnet	10.0	21.9
	Openai-o1	5.0	94.6
	DeepSeek-V3	20.0	7.3
	Qwen2.5-VL-72b-Instruct	11.0	13.9

Performance of different popular web-agent / multi-agent framework using GPT-4o as backbone on Open CaptchaWorld.

Agent Framework	MLLM Backbone	Pass@1 (%)
Human	–	93.30
Browser Use Agents	GPT-4o	5.7
SeeAct	GPT-4o	7.0
WebVoyager	GPT-4o	9.0
OWL	GPT-4o	10.0

We also measure the cost-performance trade-off among the agents:



Openai-o3 is most accurate but costly; **Gemini-2.5-Pro** and **DeepSeek-V3** offer better cost effectiveness at lower accuracy.

Success case of Agents:

Successful runs show **goal tracking + state monitoring** (e.g., iterating images until the reference matches, then submitting).

Failure case of Agents:

Reasons: Over-segmentation / overthinking & Interface & misunderstandings

Takeaway: closing the **human–agent gap** requires agents with human-like abstraction and interaction efficiency.