CUAAudit: Meta-Evaluation of Vision-Language Models as Auditors of Autonomous Computer-Use Agents

Marta Sumyk [1] Oleksandr Kosovan [1]

[1]Ukrainian Catholic University

Motivation & Research Questions

Computer-Use Agents (CUAs) are autonomous systems that execute tasks in graphical user interfaces from high-level natural-language instructions. They can click, type, scroll, and navigate desktop applications without task-specific APIs.

As CUAs become more capable, evaluation becomes a bottleneck. Existing pipelines often depend on brittle rule-based checks, benchmark scripts, or manual inspection. These methods are difficult to scale and frequently fail to capture real-world ambiguity.

This work studies Vision-Language Models (VLMs) as autonomous auditors: instead of checking internal agent logs, the auditor receives the task instruction and the final GUI state, then decides whether the task is completed.

Goal: measure whether VLM auditors are reliable enough for practical evaluation of autonomous computer-use systems. Specifically, we study the following research questions:

- **RQ1** Can VLMs accurately judge whether a CUA completed a task and well calibrated are their confidence estimates?
- **RQ2:** Do different auditors agree with each other?
- **RQ3:** How much does auditor performance depend on the operating system and corresponding benchmark?

Auditing Setup

For each task instance i, the auditor observes:

$$(x_i, d_i),$$

where x_i is the final screenshot and d_i is the natural-language task description.
The auditor predicts:

$$p_i^{(m)} \in [0, 1],$$

the confidence that the task is done, and a binary label:

$$\hat{y}_i^{(m)} \in \{0, 1\}.$$

Ground truth is the benchmark-provided binary task outcome:

$$y_i \in \{0, 1\}.$$

Auditors evaluated

- Proprietary: GPT-4o, Claude 3.5 Sonnet
- Open-source: InternVL-2-8B, LLaVA-v1.5-7B, Qwen2-VL-7B

Evaluation is not just a reporting tool. In deployment, auditor outputs may affect whether a system:

- requests human confirmation,
- accepts a task as completed,
- triggers fallback behavior,
- abstains due to uncertainty.

So an auditor must be not only accurate, but also well calibrated and consistent.

Benchmarks

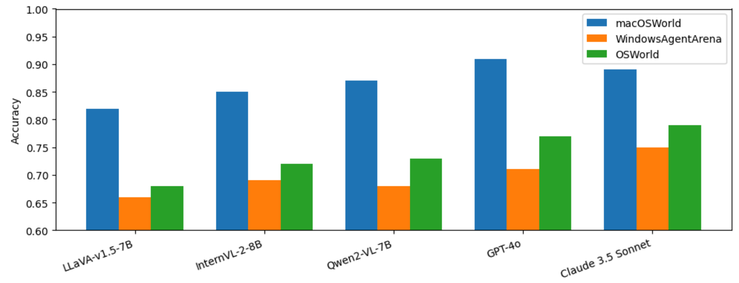
We evaluate auditors across three widely used CUA benchmarks covering three desktop operating systems:

- **macOSWorld**
- **Windows Agent Arena**
- **OSWorld** (Linux)

Together, they cover diverse tasks, applications, and interface structures. All three provide a binary *done / not done* task outcome used as ground truth.

Accuracy Across Benchmarks

Observation. All auditors perform best on macOSWorld. Performance drops on Windows Agent Arena and OSWorld, suggesting that evaluation difficulty depends strongly on environment complexity and heterogeneity.



Evaluation Metrics

1. Accuracy

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i^{(m)} = y_i\right]$$

2. Calibration via Brier Score

$$\text{Brier}_m = \frac{1}{N} \sum_{i=1}^{N} \left(p_i^{(m)} - y_i\right)^2$$

$$\text{Std}_m = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\left(p_i^{(m)} - y_i\right)^2 - \text{Brier}_m\right)^2}$$

Lower is better.

3. Inter-model Agreement

For each model pair, agreement is measured using Cohen's κ:

$$\kappa = \frac{p_o - p_e}{1}$$

Calibration Results (Brier Score, lower is better)

Auditor	macOS	Win	Linux
Proprietary			
GPT-4o	**0.058**	**0.091**	**0.074**
Claude 3.5 Sonnet	0.063	0.099	0.081
Open-source			
InternVL-2-8B	**0.097**	**0.142**	**0.118**
LLaVA-v1.5-7B	0.112	0.159	0.134
Qwen2-VL-7B	0.105	0.167	0.141

Observation. Proprietary models are substantially better calibrated. Calibration does *not* always track accuracy, so a model may be correct often while still expressing unreliable confidence.

Inter-Model Agreement

A	B	macOS	Win	Linux
GPT-4o	Claude 3.5	**0.76**	**0.66**	**0.71**
GPT-4o	InternVL	0.64	0.57	0.61
GPT-4o	LLaVA	0.61	0.54	0.59
GPT-4o	Qwen2-VL	0.66	0.58	0.63
Claude 3.5	InternVL	0.67	0.59	0.64
Claude 3.5	LLaVA	0.63	0.56	**0.66**
Claude 3.5	Qwen2-VL	**0.69**	**0.61**	0.60

Observation. Agreement is highest between proprietary auditors. Agreement drops on Windows and Linux, indicating that harder environments amplify ambiguity and subjective differences between models.

Discussion & Conclusion

The results show that VLM-based auditing is feasible, but auditor outputs should be treated as uncertain signals, not definitive judgments.

Two important implications:

- **Calibration matters.** Accuracy alone is not enough for safe deployment.
- **Disagreement matters.** If strong models disagree, the task may be ambiguous or insufficiently observable from the final GUI state alone.

Many CUA tasks depend on hidden state, background effects, or intermediate transitions that do not appear in a single screenshot.

Conclusion.

- VLM auditors can assess CUA task completion at scale.
- Proprietary models currently provide the strongest accuracy and calibration.
- Auditor performance is strongly environment-dependent.
- Inter-model disagreement reveals important ambiguity in GUI-based evaluation.

Main message: Evaluation itself is a central bottleneck for dependable deployment of Computer-Use Agents. Future auditing pipelines should explicitly model uncertainty, calibration, and evaluator variance rather than relying on binary correctness alone.

HEAL @ CHI 2026 Workshop: Human-centered Evaluation and Auditing of Language Models