“ARE WE DONE YET?”: A VISION-BASED JUDGE FOR AUTONOMOUS TASK COMPLETION OF COMPUTER USE AGENTS

Marta Sumyk[1], Oleksandr Kosovan[1] | [1]Ukrainian Catholic University | sumyk.pn@ucu.edu.ua | o.kosovan@ucu.edu.ua

Abstract

Computer Use Agents (CUAs) are designed to autonomously operate digital interfaces, yet they often fail to reliably determine whether a given task has been completed. We present an autonomous evaluation and feedback framework that uses vision-language models to assess task completion directly from screenshots and task descriptions. Our dataset covers 42 built-in macOS applications and 1,260 human-labeled tasks across a wide range of scenarios. Our framework achieves up to 73% accuracy in task success detection and yields an average relative improvement of 27% in overall task success when evaluator feedback is applied. These results show that vision-based evaluation can serve as an effective feedback mechanism that improves the reliability and self-correction of autonomous computer-use agents.

Methodology

We propose a zero-shot method using Vision-Language Models (VLMs) to automatically evaluate task completion by Computer-Using Agents (CUAs). Our pipeline operates in three steps: (1) the CUA attempts the task, (2) a VLM receives the final screenshot and task description to predict success with natural language justification, and (3) if unsuccessful, the VLM's reasoning feeds back to the agent, which repeats from its current state rather than restarting (see the diagram at the bottom). This feedback loop enables dynamic adaptation, reducing failures and redundant actions.

We evaluate three leading CUAs (Claude Computer Use, OpenAI Operator, and UI-TARS) on macOS tasks, recording full trajectories including screenshots, actions, and reasoning. Five VLMs serve as evaluators: proprietary models (GPT-4o, Claude 3.5 Sonnet) and open-source alternatives (LLaVA-v1.5-7B, InternVL 2-8B, Qwen2-VL-7B). This independent evaluation setup judges success from observable interface states, enabling robust handling of diverse applications while allowing agents to interpret feedback and adjust strategies for higher success rates.

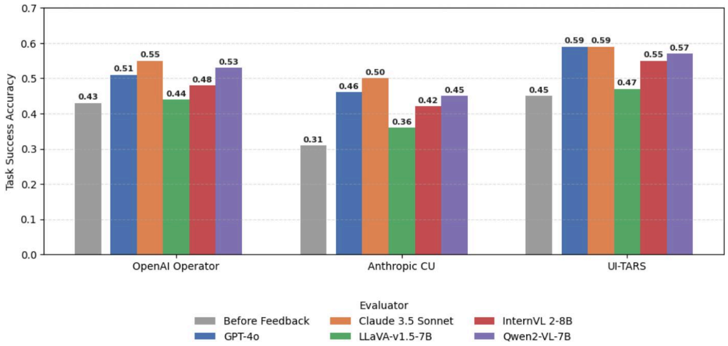
Results

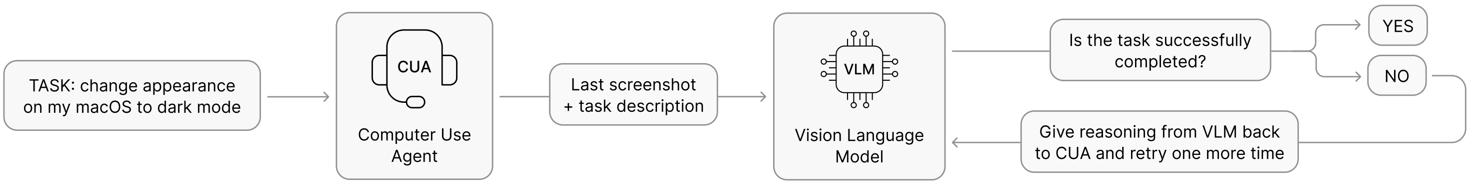
The results in Table 1 show that accuracy of task completion classification, measured against human-annotated ground truth, is consistently high for both proprietary and open-source evaluators. Even in a zero-shot setting, most models demonstrate strong alignment with human judgments, confirming that vision-language models can reliably assess task success.

Evaluator Model	OpenAI Operator	Anthropic CU	UI-TARS
Proprietary Evaluators			
GPT-4o	0.61	0.69	0.64
Claude 3.5 Sonnet	**0.69**	**0.71**	**0.73**
Open-Source Evaluators			
LLaVA-v1.5-7B	0.56	0.61	0.52
InternVL 2-8B	0.62	**0.67**	0.61
Qwen2-VL-7B	**0.68**	0.66	**0.70**

Table 1. Task completion classification accuracy (*done*/*not done*) across proprietary and open-source VLM-based evaluators for three CUAs. **Claude 3.5 Sonnet** achieves the highest proprietary performance, while **Qwen2-VL-7B** leads among open-source models.

Figure below illustrates the effect of evaluator feedback on task success rate across the three CUAs. All evaluated VLM feedback mechanisms lead to measurable performance gains compared to the baseline without feedback. These findings highlight that VLM evaluators not only reliably assess task completion but also enhance the self-correction ability of CUAs through interpretable, vision-grounded feedback.





TASK: change appearance on my macOS to dark mode

CUA

Computer Use Agent

Last screenshot + task description

VLM

Vision Language Model

Is the task successfully completed?

YES

NO

Give reasoning from VLM back to CUA and retry one more time