AI Agents Finished the Experiments but Failed the Research

Kabui, Charles

2026-07-31

Read at ToKnow.ai

Frontier AI agents can run experiments and write papers, but that did not make them capable researchers in a new shadow evaluation. Researchers gave agents the central questions from two unpublished NeurIPS 2026 submissions, then asked the original authors to review the agents' independent work. Each main run received six days, $3,000 in Anthropic API credits, GPU access, a virtual machine, and the open web. The agents completed the engineering

without human help, yet both papers were clear rejections. Reviewers scored them 2/6 and 1/6 overall. A repeat using GPT-5.6 Sol with its native Codex scaffold showed nearly the same failures, which makes a single bad model or scaffold a less likely explanation.

The agents recognized promising directions but tested them on small synthetic or hand-picked datasets, treated weak negative results as findings, and failed to change course after poor reviews. Both main runs spent less than half their API budgets and retired their most ambitious goals within ten hours. This matters because coding benchmarks test whether an agent can reach a fixed answer. Research requires deciding what evidence counts, noticing when an approach is going nowhere, and trying something genuinely different. The released logs make that gap inspectable rather than hypothetical. Two studies cannot settle what all research agents can do, but they show why successful engineering alone is weak evidence for automated discovery.

Read More: An AI System That Writes Medical Research Papers Good Enough for Peer Review

Sources:

- Shadow Evaluation Paper
- CRUX Research Evaluation
- Original Authors' Expert Reviews
- Released Agent Logs
