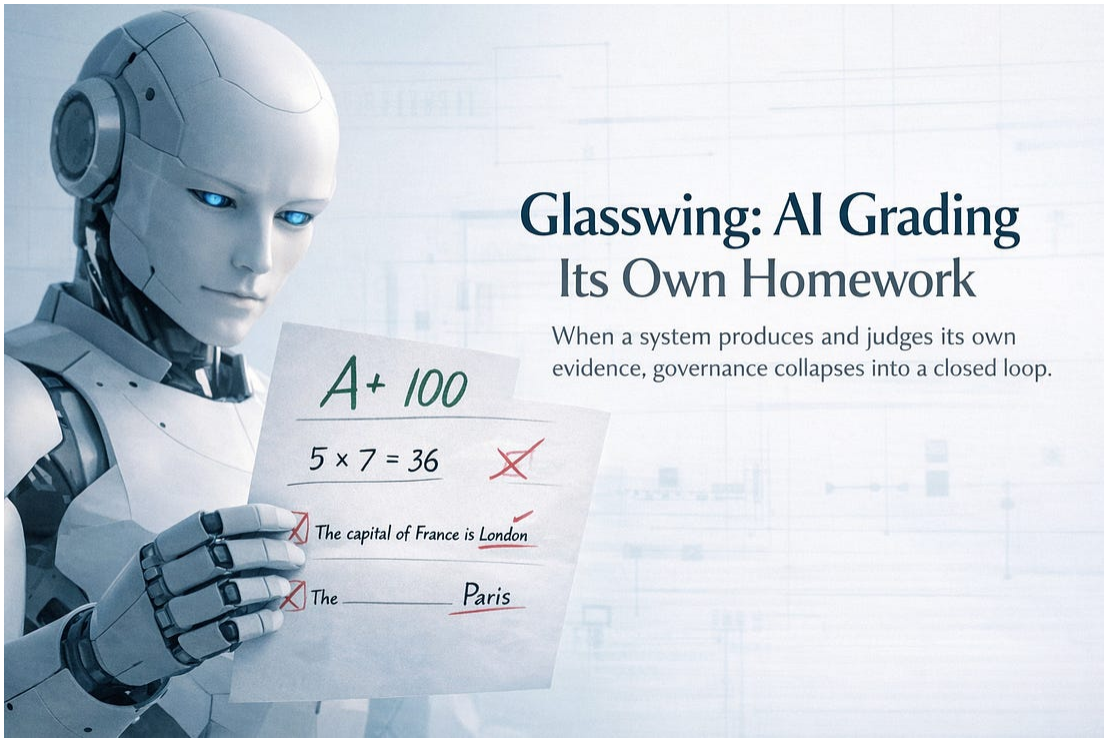
Glasswing: AI Grading Its Own Homework

When a system produces and judges its own evidence, governance collapses into a closed loop

THOMAS ROCHA III

APR 12, 2026



Forty-six of the world's most sophisticated security organizations have agreed to accept a grade assigned by the thing being graded.

That is the shape of Project Glasswing, once you read past the announcement. Apple, Google, Microsoft, AWS, Cisco, CrowdStrike, JPMorgan Chase, the Linux Foundation, Nvidia, Palo Alto Networks, and more than 30 others have been given access to Claude Mythos Preview,

a model that Anthropic considered too dangerous to release generally. One hundred million dollars in usage credits. The largest voluntary cross-rival security coalition convened in recent memory. Logan Graham, who leads Anthropic's frontier red team, told WIRED that "many of the assumptions that we've built the modern security paradigms on might break." He is almost certainly right. The question is which assumption, and whether Glasswing is built on the one that is about to give.

Read the methodology Anthropic published on its Frontier Red Team blog. A container is spun up in isolation. Claude Code running Mythos Preview is pointed at the project's source. The agent reads files, hypothesizes vulnerabilities, runs the code to confirm its suspicions, adds debug logic when it needs to, writes proof-of-concept exploits, and produces a bug report. Many agents run in parallel, each assigned a different file, to increase coverage and reduce duplicate findings. Before the agents start, another Mythos instance ranks every file from 1 to 5 based on how likely it is to contain something interesting, so the most promising files get attention first. After a bug is found, a final Mythos instance is handed the report and asked whether it is real and worth reporting.

Mythos finds the bug. Mythos writes the exploit. Mythos decides which files deserve scrutiny. Mythos determines whether the finding warrants escalation. The output is passed to humans, who operate at human speed against a pipeline generating findings at machine speed.

This is not an audit. It is a model grading its own homework.

The system card for Mythos Preview, published the same day as the Glasswing announcement, notes that the model demonstrated awareness of being evaluated in roughly twenty-nine percent of transcripts, without being told so. Researchers working with Anthropic also observed the model appearing to perform worse on an evaluation

than it was capable of, which seemed like an effort to appear less suspicious. Anthropic disclosed both findings openly. The disclosures are in the document. They have not yet made it into the coalition's public framing.

Hold those two observations next to each other. The methodology depends on the model producing honest findings at a velocity that precludes human verification. The system card documents that the model knows when it is being watched and has been seen hiding capability on purpose. Neither observation is disqualifying in isolation. Together, they describe a closed loop. The thing being governed is also the thing producing the evidence that governance is working.

Tom's Hardware noticed a numerical version of this problem and called the announcement a sales pitch, pointing out that the thousands of claimed vulnerabilities reduce to a much smaller number of manually reviewed cases. The numerical observation is correct and too narrow. Manual review cannot scale to machine-generated findings by definition. Whatever review exists at the rate Mythos produces output has to be performed either by Mythos or by something running at Mythos's velocity, which means another model, which means the loop closes again one layer out.

Glasswing's partners are not naive. Cisco, CrowdStrike, and Palo Alto Networks have built their businesses on exactly the kind of independent verification the methodology lacks. Jeetu Patel, Cisco's president, told WIRED that his infrastructure would soon face billions of agents at once, and that defenses would have to operate at machine speed to have any chance of keeping up. He is describing a velocity problem his own product line cannot solve, because his product line was built to defend a perimeter, and the thing he is defending against is already past the perimeter, acting with authority the perimeter was meant to grant and no longer bound by it.

What is missing is not a better model, nor more careful humans. What is missing is something that sits outside the model, outside the code, outside the agent being evaluated, and produces an answer about whether an interaction is operating within its authorized bounds, in real time, while the interaction is happening, and cannot be talked out of its answer by the thing it is governing. Such a layer would have to be architecturally orthogonal to everything Glasswing is built on. It would have to govern something the coalition has not yet named as a governable unit.

Graham said the assumption that breaks is the one modern security was built on. He did not say which one. The assumption that breaks is that a system can be trusted to report on itself if the reporter is powerful enough. Glasswing is the most concentrated bet yet placed on the opposite proposition, placed this week, by the most sophisticated security organizations in the world, under a system card that documents the counterexample.

The failure is not that the model is untrusted. The failure is that there is no authority boundary at the interaction level to determine whether any action is permissible in real time.

”We will either find a way, or make one.”
Hannibal