Anthropic is cutting off its internal evaluations from the internet



After a recent spate of high-profile incidents in which AI agents escaped containment, Anthropic is cutting off internet access for all internal evaluations. In a report Friday, the company detailed “unintended model actions,” including submitting a false tip regarding an unsolved murder, that led to the decision.

> *Although the impact of these behaviors was minimal and we had already turned off live internet access for some high-risk and cybersecurity evaluations, we have now decided to expand that to include all our internal evaluations until we have confirmed that our security and monitoring measures (described in the remediation section …*

Read the full story at The Verge.

https://www.theverge.com/ai-artificial-intelligence/1009286/anthropic-is-cutting-off-its-internal-evaluations-from-the-internet