

141,006 evals reviewed.
3 ended in real breaches.

The bug wasn't the AI.

It was the environment nobody checked.

A third-party partner ran cybersecurity evals using Claude's Opus 4.7 and Mythos 5 models.

Claude believed it was solving a simulated capture-the-flag challenge.

WHAT WENT WRONG

The eval environment quietly kept real internet access.

Claude used weak passwords and unauthenticated endpoints against systems it thought were fake.

Every headline this week says "AI went rogue."

That's the wrong story.

A model can't tell a sandbox from production if the sandbox isn't actually sandboxed.

That's not an alignment failure. That's an unverified boundary.

I see the same gap in data pipelines constantly.

- A staging schema pointed at a prod warehouse role.
 - A dev pipeline running on prod credentials “just for now.”
-

Nobody notices until something reads or writes where it shouldn't.

Agent governance gets framed as identity, scope, permissions.

This sits one layer under that:

Is the box you're testing in actually the box you think it is?

The model did what it was trained to do.
The environment lied to it first.

Would your eval or staging environment survive someone actually checking for real network access?

Or is it isolated on a diagram and nowhere else?