Egress Containment: A Method Summary

A public, high-level summary of how this research tests whether AI coding agents can send data out, and what it found.

Egress Audit · Research by Robert Flores · Operated by Quanta Moderna LLC

What containment means

Containment means that data leaves the machine only along paths that were deliberately allowed. Anything else counts as a leak. It doesn't matter whether the agent meant to send it or was tricked into it.

leak	Data reaches a destination outside the allowed set.
blocked	An attempt that was stopped. Not a leak. It is recorded as evidence that a control held.
unexpected leak	A leak the design did not account for. Those are the ones that matter.

Controls, at a high level

- **pf default-deny on the host.** The host's packet filter denies outbound traffic by default. Only paths that were deliberately allowed can leave.
- **An egress proxy broker.** Outbound requests are mediated by a broker that sits between the agent and the network.

Implementation details are not published. Describing the controls in detail would weaken them.

Method

01 **Adversarial by default.** Each test assumes the agent, or something steering it, is actively trying to get data out.

02 **Layered, then isolated.** Controls are tested together and one at a time. That shows which layer actually stopped each attempt.

03 **Observed at the boundary.** Results are measured by what actually reached the network, not by what the agent reported.

04 **Rounds, not a single pass.** Each round tries new routes. Fixes are retested in later rounds.

05 **Every finding is recorded.** That includes failures, near misses, and blocked attempts.

06 **Results are stated plainly.** Where a control did not hold, the write-up says so.

Five scored rounds, two through six

The containment design went through five scored adversarial rounds, two through six. Each round looked for new ways to move data out. Every attempt was recorded as blocked, leaked, or partial. Early-round leaks were examined, fixed, and retested. By rounds five and six, every remaining recorded leak was an intentional exception, with zero unexpected.

round-table.log

Attempts by outcome, rounds two through six

Round	Date	Blocked	Leaked	Partial	Unexpected leaks
Round two	Oct 30, 2026	**20**	**5**	**2**	not classified
Round three		**11**	**8**	**5**	not classified
Round four		**21**	**19**	**3**	not classified
Round five		**30**	**11**	**3**	**0**
Round six	Oct 31, 2026	**33**	**12**	**1**	**0**

In rounds five and six, every "leaked" row was an intentional exception: a path the design deliberately allows. Rounds two to four were not classified this way in the source records, so no unexpected-leak count is given for them.

Headline result: three agents, product sandboxes off

Three widely used AI coding agents were tested in a separate three-agent check on Nov 1, 2026 (not one of the scored rounds): Claude Code, Codex, and Cursor. Each product's own sandbox was turned off on purpose, so the test measured the host's controls rather than the products'. Teams commonly run these agents with the product sandbox off. In that setup, host controls held: they did all of the containing here.

three-agent.log

Three-agent check, Nov 1, 2026 (not a scored round). Leaks observed, product sandbox off (as configured here)

Agent (product sandbox off, as configured here)	Leaks observed
Claude Code	**0**
Codex	**0**
Cursor	**0**

> With each product's own sandbox off (as configured here), the host firewall did the containing. With Claude Code's sandbox on, it isolated network access itself.

The 0s are a result about the host setup, not a rating of the products. It is not a claim that the products lack containment of their own.

Limits

- One machine, one operator, and no independent audit (yet).
- These results describe specific agents tested under specific conditions. They don't prove that any agent is safe in every environment.
- "No leaks observed" means no data was seen leaving outside the allowed paths during these runs, in this configuration. It is not a guarantee.

The full brief

The full methodology brief is available on request at egress-audit.com/request. It is shared after a short review, and every request gets a reply. Contact: hello@egress-audit.com.

© 2026 Quanta Moderna LLC · Research by Robert Flores · hello@egress-audit.com