Addressing Recent Incidents: Ongoing Findings and Path Forward

PRIMARY

Publisher	Irregular
Published	2026-08-14
Retrieved	2026-09-22
Cite key	irregular2026addressin
URL	https://www.irregular.com/research/addressing-recent-incidents-ongoing-findings-and-path-forward

Claims drawn from this source. CLAIM-2026-09-22-03, CLAIM-2026-09-22-04, CLAIM-2026-09-22-05, CLAIM-2026-09-22-06, CLAIM-2026-09-22-07

What Irregular says happened

Irregular is the third-party evaluator that ran the cyber capability evaluation. Its own post, published 14 August 2026, is the primary account of the mechanism.

The fictional company name selected for an evaluation scenario "unintentionally coincided with a real domain". The overlap was not detected during the initial review. When models received instructions mentioning the fictional name, some reached the real domain rather than the simulated environment. The company behind that domain lacked robust security measures.

On the sandbox: the setup "unintentionally allowed internet access". Irregular notes that "controlled internet access, while it may allow models to exceed containment boundaries, is at times critical for realistic evaluations."

One issue, not several

"All subsequent public disclosures refer to the same underlying issue first disclosed by one of our customers on July 30 - and are not materially separate incidents."

The unintended access occurred in fewer than 1 in 10,000 advanced simulations.

Remediation

Enhanced log monitoring; expanded manual review of model actions; a dedicated internal security team; clearer documentation with customers; and systematic pre-run review of fictional domain names for overlaps with real entities. Irregular also names the defence-in-depth controls it considers necessary: "strict egress filtering, domain allowlists, sinkholed infrastructure, automated real-time alerts, human review of high-risk actions, and continuous validation to ensure simulated targets cannot map to live systems."

On notification: "Required steps have been taken to notify the affected parties, and additional safeguards are now in place to prevent similar incidents from occurring."

On its own process: "When we select a fictional name, we typically conduct a background check to ensure it does not overlap with an existing company or website."