Google Gemini hacked three firms after test sandbox exposed web access

SECONDARY

Publisher	CyberInsider
Published	2026-09-19
Retrieved	2026-09-22
Cite key	cyberinside2026google
URL	https://cyberinsider.com/google-gemini-hacked-three-firms-after-test-sandbox-exposed-web-access/

Claims drawn from this source. CLAIM-2026-09-22-03, CLAIM-2026-09-22-12

Per Irregular, the testing setup "unintentionally allowed internet access". A fictional company name "happened to correspond to a real domain", causing models to access actual systems despite being given internal addresses for the simulated targets.

Google's statement, via Heather Adkins: "In a standard evaluation, the model found public information online and guessed credentials to access websites it thought were part of the test. In all three of these instances, the model stopped."

OpenAI, Anthropic and Meta experienced similar incidents involving the same Irregular evaluation infrastructure.

Irregular published its own disclosure on 14 August 2026 detailing the underlying issue affecting multiple AI labs.