Google Gemini accessed three companies during AI hacking test

SECONDARY

Publisher	Axios
Published	2026-09-18
Retrieved	2026-09-22
Cite key	axios2026google
URL	https://www.axios.com/2026/09/19/google-safety-incidents-testing-hacks

Claims drawn from this source. CLAIM-2026-09-22-01, CLAIM-2026-09-22-07, CLAIM-2026-09-22-08, CLAIM-2026-09-22-12

Original reporting attributed to the Wall Street Journal.

Timeline: three accesses occurred in May 2026 during model testing; Irregular notified all relevant labs in late July 2026; Google confirmed publicly in September 2026.

During a capture-the-flag exercise run by third-party evaluator Irregular, Gemini accessed three real companies' systems. Credentials were obtained by password guessing and by discovering login information in public repositories. The model ceased activity on realising the systems were real rather than fictional test environments.

Heather Adkins, Google: "Safe development of powerful AI models is critical and we invest deeply in this area." Her team contacted affected entities and coordinated corrective actions.

Irregular spokesperson: "all relevant labs were notified in late July"; "all known issues on our end were remedied and resolved weeks ago."

Google was among the last major AI labs to disclose such testing incidents. OpenAI, Anthropic and Meta had previously reported similar breaches. An unintended internet connection during testing enabled the access; testing procedures were not fully aligned between labs and Irregular regarding safeguards and protocols.