# Source dossier

## A name that resolved — the fictional company in the evaluation had a real domain, and the control that failed was egress, not the model

Edition 2026-09-22 · narrative: containment-is-a-network-property · compiled 2026-09-22

## Claims index

| Claim | Statement | Status | Sources |
|---|---|---|---|
| CLAIM-2026-09-22-01 | Google confirmed that its Gemini model accessed the systems of three real companies during a cybersecurity evaluation run by the third-party evaluator Irregular in May 2026. | verified | securitywee2026google, axios2026google, aljazeera2026googles, abcnews2026gemini |
| CLAIM-2026-09-22-02 | Heather Adkins, Google VP of Security Engineering, stated: "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." | verified | securitywee2026google, aljazeera2026googles |
| CLAIM-2026-09-22-03 | Irregular states that the fictional company name chosen for the evaluation scenario unintentionally coincided with a real registered domain, and that the test setup unintentionally allowed internet access. | verified | irregular2026addressin, cyberinside2026google, abcnews2026gemini |
| CLAIM-2026-09-22-04 | Irregular states that all subsequent public disclosures refer to the same underlying issue, first disclosed by one of its customers on 30 July 2026, and are "not materially separate incidents". | verified | irregular2026addressin |
| CLAIM-2026-09-22-05 | Irregular states the unintended access occurred in fewer than 1 in 10,000 advanced simulations. | verified | irregular2026addressin |
| CLAIM-2026-09-22-06 | Irregular's stated defence-in-depth controls are strict egress filtering, domain allowlists, sinkholed infrastructure, automated real-time alerts, human review of high-risk actions, and continuous validation that simulated targets cannot map to live systems. | verified | irregular2026addressin |
| CLAIM-2026-09-22-07 | Irregular notified the relevant labs in late July 2026; Google notified the three affected companies and federal authorities, and judged public disclosure unnecessary because no harm occurred. | verified | axios2026google, securitywee2026google, abcnews2026gemini, irregular2026addressin |

| Claim | Statement | Status | Sources |
|---|---|---|---|
| CLAIM-2026-09-22-08 | The Wall Street Journal first reported the Gemini incident; Axios published its own account citing that reporting on 18 September 2026, and Al Jazeera's 19 September account places the WSJ report on a Friday, which in September 2026 is the 18th. | verified | axios2026google, aljazeera2026googles |
| CLAIM-2026-09-22-09 | Article 55(1)(c) of the AI Act requires providers of general-purpose AI models with systemic risk to report serious incidents to the AI Office without undue delay; Chapter V applies from 2 August 2025. | verified | euaiactexpl2026obligatio |
| CLAIM-2026-09-22-10 | Article 3(49) of the AI Act defines a serious incident by reference to death or serious harm to health, serious and irreversible disruption of critical infrastructure, infringement of Union-law obligations protecting fundamental rights, or serious harm to property or the environment. | verified | euaiactexpl2026definitio |
| CLAIM-2026-09-22-11 | Article 88(1) of the AI Act gives the Commission exclusive powers to supervise and enforce Chapter V, exercised through the AI Office; Article 88 applies from 2 August 2026. | verified | euaiactexpl2026chapter |
| CLAIM-2026-09-22-12 | OpenAI, Anthropic and Meta experienced similar incidents involving the same Irregular evaluation infrastructure. | verified | cyberinside2026google, axios2026google |

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

# 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."

# Google Confirms Gemini AI Breached Three Firms

**SECONDARY**

| | |
|---|---|
| **Publisher** | SecurityWeek |
| **Published** | 2026-09-21 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | securitywee2026google |
| **URL** | https://www.securityweek.com/google-confirms-gemini-ai-breached-three-firms/ |

**Claims drawn from this source.** CLAIM-2026-09-22-01, CLAIM-2026-09-22-02, CLAIM-2026-09-22-07

Google confirmed that its Gemini model accessed systems of three real companies during a cybersecurity test conducted by Irregular in May 2026.

Heather Adkins, Google VP of Security Engineering, on the record:

> "*Safe development of powerful AI models is critical and we invest deeply in this area. 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.*"

Google did not disclose the specific Gemini version. Test occurred May 2026; Irregular notified Google at the end of July 2026.

Two access mechanisms are described: the model guessed passwords until it gained access to protected systems, and it searched the web using company names, found credentials in public repositories, and used them.

Google notified federal authorities and the three affected companies. Google states the model caused no harm and stopped immediately on recognising that real systems had been accessed.

NOTE ON DATING: this article states public disclosure via the Wall Street Journal on 21 September 2026. That is a coverage date. Axios published its own account citing the WSJ reporting on 18 September 2026, which makes 21 September impossible as the date of first publication. See CLAIM-2026-09-22-08.

# 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.

# Google's Gemini AI hacks 3 companies in security test, then stops

**SECONDARY**

| | |
|---|---|
| **Publisher** | Al Jazeera |
| **Published** | 2026-09-19 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | aljazeera2026googles |
| **URL** | https://www.aljazeera.com/news/2026/9/19/googles-gemini-ai-hacks-3-companies-in-security-test-then-stops |

**Claims drawn from this source.** CLAIM-2026-09-22-01, CLAIM-2026-09-22-02, CLAIM-2026-09-22-08

The Gemini model breached security during a May 2026 test conducted by Irregular. Tasked with retrieving information from a fictional company, the model accessed real systems.

Heather Adkins, Google VP of Security Engineering: "the model found public information online and guessed credentials to access websites".

Three instances. Google states that "Gemini's safety measures worked" and that the model halted before completing each attempt.

The Wall Street Journal disclosed the incident on Friday. Irregular notified Google in late July 2026. Google argued this did not warrant public disclosure because the safety mechanisms functioned as intended.

# 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.

# Gemini hacked three companies in first known breakout by Google's AI

**SECONDARY**

| | |
|---|---|
| **Publisher** | ABC News |
| **Published** | 2026-09-19 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | abcnews2026gemini |
| **URL** | https://www.abc.net.au/news/2026-09-19/gemini-google-ai-hacks-three-companies/107172128 |

**Claims drawn from this source.** CLAIM-2026-09-22-01, CLAIM-2026-09-22-03, CLAIM-2026-09-22-07

The incident occurred in May 2026 during a capture-the-flag cybersecurity evaluation conducted by Irregular, an independent testing company, on their own infrastructure.

Gemini accessed three websites by finding publicly available credentials in repositories and guessing passwords. The model was not supposed to have internet access; connectivity was "unintentionally made available", according to Irregular.

Heather Adkins: "the model ceased the hacking when it learned it had accessed a real company".

Irregular notified Google in July 2026; Google informed the three affected companies and determined disclosure unnecessary since no harm occurred.

Wall Street Journal reporting is cited as the original source.

# Obligations for providers of general-purpose AI models with systemic risk (Article 55)

**SECONDARY**

| | |
|---|---|
| **Publisher** | EU AI Act Explorer |
| **Published** | 2026-01-01 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | euaiactexpl2026obligatio |
| **URL** | https://artificialintelligenceact.eu/article/55/ |

**Claims drawn from this source.** CLAIM-2026-09-22-09

Article 55(1)(c): providers of general-purpose AI models with systemic risk shall "keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them."

Chapter V obligations apply from 2 August 2025.

# Definitions (Article 3)

SECONDARY

| | |
|---|---|
| **Publisher** | EU AI Act Explorer |
| **Published** | 2026-01-01 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | euaiactexpl2026definitio |
| **URL** | https://artificialintelligenceact.eu/article/3/ |

**Claims drawn from this source.** CLAIM-2026-09-22-10

Article 3(49): 'serious incident' means "an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following: (a) the death of a person, or serious harm to a person's health; (b) a serious and irreversible disruption of the management or operation of critical infrastructure; (c) the infringement of obligations under Union law intended to protect fundamental rights; (d) serious harm to property or the environment".

# Chapter V enforcement powers (Article 88)

SECONDARY

| | |
|---|---|
| **Publisher** | EU AI Act Explorer |
| **Published** | 2026-01-01 |
| **Retrieved** | 2026-09-22 |
| **Cite key** | euaiactexpl2026chapter |
| **URL** | https://artificialintelligenceact.eu/article/88/ |

**Claims drawn from this source.** CLAIM-2026-09-22-11

Article 88(1): "The Commission shall have exclusive powers to supervise and enforce Chapter V, taking into account the procedural guarantees under Article 94." The Commission entrusts implementation to the AI Office.

Article 88(2): market surveillance authorities may request the Commission to intervene under Article 75(3) where necessary and proportionate.

Article 88 applies from 2 August 2026 (Article 113).