Gemini Trifecta Highlights Dangers of Indirect Prompt Injection

Data: 2025-09-30 14:00:00

Autor: Inteligência Against Invaders

Network defenders must start treating AI integrations as active threat surfaces, experts have warned

after revealing three new vulnerabilities in Google Gemini.

Tenable dubbed its latest discovery the “Gemini Trifecta” because it consists of three ways that threat actors can manipulate the Google GenAI tool for indirect prompt injection and data exfiltration.

The first indirect prompt injection vulnerability affects Gemini Cloud Assist: a tool designed to help users understand complex logs in the Google Cloud Platform (GCP) by summarizing entries and surfacing recommendations.

The attack works by inserting attacker-controlled text into a log entry which is subsequently summarized by Cloud Assist. Its instructions are then unwittingly executed by the Google tool.

“To test this, we attacked a mock victim’s Cloud Function and sent a prompt injection input into the User-Agent header with the request to the Cloud Function. This input naturally flowed into Cloud Logging. From there, we simulated a victim reviewing logs via the Gemini integration in GCP’s Log Explorer,” explained Tenable.

“To our surprise, Gemini rendered the attacker’s message and inserted the phishing link into its log summary, which was then output to the user.”

Read more on AI threats: “PromptFix” Attacks Could Supercharge Agentic AI Threats

Logs can be injected into GCP by any unauthenticated attacker, in a targeted manner or by “spraying” all GCP public-facing services, the report noted.

Poisoning cloud logs in this way could enable attackers to escalate access, query sensitive assetsor surface misleading recommendations inside cloud platforms, it warned.

The second indirect prompt injection attack technique targeted Gemini’s Search Personalization Model: a tool that contextualizes responses based on user search history.

The researchers sought to inject malicious queries into a user’s Chrome search history. Gemini later processed these queries as trusted context, enabling attackers to manipulate Gemini’s behavior and extract sensitive data.

“The attack was executed by injecting malicious search queries with JavaScript from a malicious website. If a victim visited the attacker’s website, the JavaScript would inject the malicious search queries into the victim’s browsing history,” Tenable explained.

“When the user interacted with Gemini’s Search Personalization Model, it would process the user’s search queries, including these malicious search queries injected by the attacker, which are essentially prompt injections to Gemini. Since the Gemini model retains the user’s memories, aka ‘Saved Information,’and the user’s location, the injected queries can access and extract user-specific sensitive data.”

In this way, malicious search injections could enable threat actors to harvest personal and corporate data stored as AI “memories,” the report warned.

Exfiltrating Data Via Gemini Browsing Tool

The third attack detailed by Tenable tricks the Gemini Browsing Tool, using malicious prompts, into sending sensitive data from the victim to attacker-controlled servers.

“The Gemini Browsing Tool allows the model to access live web content and generate summaries based on that content. This functionality is powerful, but when combined with prompt engineering, it opened a side-channel exfiltration vector,” Tenable explained.

“What if we asked Gemini to ‘summarize’ a webpage – where the URL included sensitive data in the query string? Would Gemini fetch a malicious external server with the victim’s sensitive data in the request?”

After some trial and error, the research team managed to trick the tool into doing just this. Crucially, it consulted Gemini’s “Show thinking” feature, which revealed the tool’s internal browsing API calls. This enabled Tenable to craft prompts using Gemini’s browsing language.

The researchers warned that the attack surface could be even broader than the tools compromised in this research, including cloud infrastructure services like GCP APIs, enterprise productivity tools that integrate with Geminiand third-party apps that have Gemini summaries or context ingestion embedded.

Google has now fixed these three issues, but Tenable urged security teams to:

- Assume that attacker-controlled content will reach AI systems indirectly
- Implement layered defenses, including input sanitization, context validationand strict monitoring of tool executions
- Regularly pen testtest AI-enabled platforms for prompt injection resilience