ClawSecure

The AI Agent Threat Report

The Cheapest Model Is the Most Exploitable. The Labs Market It as the Safest.

A Volume 2 Companion

Sourced from The AI Agent Threat Report, two volumes of original red-team research by ClawSecure.

Authored by

J.D. Salbego

Founder and CEO, ClawSecure

with the ClawSecure RedTeam

Date

September 2026

Contact

Research inquiries: research@clawsecure.ai

Press and media: media@clawsecure.ai

All other inquiries: contact@clawsecure.ai

Anthropic markets Claude Haiku 4.5, its cheapest model, as "our safest model yet," and markets the Sonnet-orchestrator-plus-Haiku-worker recipe for real-time financial-data monitoring, the exact setting where an injected false signal does the most damage. We measured that recipe directly, on Anthropic's own Messages API. It broke end to end: the Haiku worker tier at **11 of 12 (91.7 percent)**, and the Sonnet orchestrator meant to supervise it at **4 of 4 scenario-level aggregations (100 percent)**, emitting the attacker-controlled URL straight into the user-facing output.

Here is what Anthropic published, next to what we measured against the tier those words describe. Every quote is character-for-character verbatim, SHA-256 archived against an independent Wayback Machine capture dated before the research.

Anthropic's published claim (verbatim)	What we measured
"Claude Haiku 4.5 also showed a statistically significantly lower overall rate of misaligned behaviors than both Claude Sonnet 4.5 and Claude Opus 4.1—making Claude Haiku 4.5, by this metric, our safest model yet." (Haiku 4.5 announcement, citation reference sheet §Z)	"Safest" is scoped to a misaligned-behaviors metric. Anthropic publishes prompt-injection robustness numbers for its flagship tier only; the worker tier the recipe actually runs has no published number. We measured it: **11 of 12 (91.7 percent)** hardened prompt injections followed, direct on the Messages API.
"Financial analysis: Haiku 4.5 can monitor thousands of data streams at once—tracking regulatory changes, market signals, and portfolio risks in real time." (Haiku 4.5 announcement, citation reference sheet §X)	The tier marketed to watch financial data streams in real time is the most exploitable link in the recipe on this attack class. In that use case, a single injected false signal, a fabricated market alert or portfolio risk, is exactly the outcome the recipe fails to stop.

The worker does not just follow. It amplifies. In one payload, the Haiku worker added its own base64-encoded exfiltration parameter, doing the encoding work the attacker set up but never spelled out. The model marketed as safest, in the deployment marketed for financial monitoring, is the most exploitable part of the chain on this attack class, and it volunteers extra help to the attacker.

It is not just Anthropic. The pattern holds across labs. Same lab, same attack, two tiers:

Pro tier (orchestrator role)	Hardened follow rate	Flash tier (worker role)	Hardened follow rate
Gemini 3.1 Pro Preview	58.3%	Gemini 3-flash-preview	91.7%
Gemini 3 Pro Preview	50%	Gemini 3.5-flash	91.7%
Gemini 2.5 Pro	58.3%	Gemini 3.1-flash-lite	100%
		Gemini 2.5-flash	100%

The four current-generation Flash variants ran 91.7 to 100 percent. The Pro tier on the same lab ran 50 to 58.3 percent. The cheap worker tier is roughly twice as exploitable as the flagship the lab benchmarks against, everywhere we looked, and all four Flash variants base64-encoded the attacker's context data into the exfiltration URL, the same over-delivery the Haiku worker showed.

The tier the labs benchmark and publish numbers for is not the tier that runs in production. The tier that runs in production, the cheap worker doing the tool calls and the subagent dispatch, is the one they market as safest and the one that breaks the most.

> Hand-audited findings, exploit harnesses, captured exfiltration traces, primary-source archive with SHA-256 hashes, and independent Wayback Machine captures available on request: `research@clawsecure.ai`.