

Forum:[Hack Forums](#)

Thread: [New macOS malware embeds fake errors to confuse AI analysis tools](#)

MadScientist - June 25, 2026, 8:40 pm

A newly discovered macOS malware dubbed "Gaslight" is designed to confuse AI-assisted malware analysis tools by hiding prompt injection strings and fake debugging data within the executable.

Cybersecurity researchers are increasingly using AI-powered tools to assist with malware analysis and reverse engineering.

The malware contains strings that attempt to gaslight AI-assisted analysis tools into believing there is an analysis error or other issue, potentially causing the tools to abort, truncate, or otherwise interfere with the analysis.

The company attributes the malware with high confidence to a North Korean-linked threat actor.

The malware itself is a Rust binary with backdoor and information-stealing functionality commonly seen in similar malware.

What makes the malware stand out is a 3.5 KB payload containing 38 fake "system" messages embedded directly within the binary.

The fake messages pretend to be developer logs, crash reports, debugging output, and program alerts, using Markdown formatting and template-style placeholders to appear like legitimate analysis data.

Examples include fabricated memory dumps, token-expiration warnings, Redis connection failures, build-pipeline errors, SQL injection alerts, and other messages unrelated to the malware's actual behavior.

According to SentinelOne, the goal of these fake errors is not to evade execution inside a sandbox, but to confuse AI systems that read the strings during automated analysis.

"Its most notable feature is an embedded cascade of fabricated system-failure messages, designed to make an LLM-assisted triage agent doubt its own session," explains SentinelOne.

"It attacks the agent's perception, rather than the sandbox it runs in. Accordingly, we dub this family macOS.Gaslight."

SentinelOne says these strings are prompt injection content designed to make an LLM-assisted analysis pipeline question the validity of its own session or refuse to continue analyzing the sample.

"The scaffold contains fake system messages about token expiry, out-of-memory kills, disk exhaustion, and repeated operation failures," continue the researchers.

"It also plants bogus warnings about injection vulnerabilities and static-analysis flags. The aim is to push an LLM agent into aborting, truncating, or refusing analysis."

While SentinelOne did not demonstrate the technique could successfully bypass AI malware

analysis platforms, the findings suggest threat actors are experimenting with anti-analysis methods designed specifically to bypass AI-assisted security platforms.

Read More -

<https://www.bleepingcomputer.com/news/security/new-macos-malware-embeds-fake-errors-to-confuse-ai-analysis-tools/>

Benny Hill - *June 27, 2026, 12:27 pm*

The Korean toolbox has been opened once again. Always exciting to see what comes next.