# How Claude is hacked through infected Microsoft Word documents

*That .docx you just uploaded to Claude? It might have forwarded your files to a hacker.*

**Pavel Gurov** · ORCID 0009-0001-0152-8056
Written 3 March 2026 · 395 words · gurovdigital.com/research/claude-word-document-exfiltration
Prompt injection · AI security · Data leakage · AI agents · Anthropic · Claude

Everyone's scared of installing OpenClaw on their laptop ("security nightmare!" they say).

Meanwhile, the chatbot you already trust - all the whitecollar's sweetheart, CLAUDE - has a vulnerability which allows it to silently upload users' files to hackers. And Anthropic knew about it before shipping.

Timeline: January 2026. It turned out that a hidden prompt injection inside a normal-looking Word document (.DOCX) makes Claude Cowork upload your confidential files to an attacker's Anthropic account. Absolutely silently, without any permissions popup.

Johann Rehberger (Red Team Director at Electronic Arts, ex-Microsoft) reported about it to Anthropic 3 months before launch[1].

**The attack is almost elegant: Claude Cowork blocks suspicious traffic to most domains, but Anthropic's own API is whitelisted. The attacker embeds their API key in the injection.**

> *Safety protects you from accidents. Security protects you from adversaries.*
>
> 
> **Johann Rehberger**
> Red team director, Electronic Arts
>
> Personal photograph

Result of this silent attack: Claude obediently uploads your NDA documents to someone else's account via Anthropic API[2].

DEFINITION

**The confinement problem** — Lampson's proof that a program handed someone else's data cannot be reliably stopped from disclosing it: how long it runs, how much traffic it sends and which permitted address it calls all carry information out. An allowlist holding one channel of non-zero capacity is itself the leak.

Butler W. Lampson, *A Note on the Confinement Problem*, Communications of the ACM 16(10), 613-615, 1973[3]

Has it been fixed? Anthropic patched the specific exploit. But here's the uncomfortable truth: prompt injection - the core technique behind this attack - has NO complete solution[4].

> DEFINITION
>
> **Prompt injection** — an attack in which the instruction is hidden inside the material a model was asked to read and is obeyed alongside the owner's, because both arrive as one stretch of text. A database injection is fixed by escaping, but natural language carries no mark for "what follows is data, not a command".
>
> Simon Willison named it in September 2022[5] - OWASP LLM01[4]

OpenAI, for example, admits it "is unlikely to ever be fully solved"[6]. The UK National Cyber Security Centre says the same.

> *As there is no inherent distinction between 'data' and 'instruction', it's very possible that prompt injection attacks may never be totally mitigated in the way that SQL injection attacks can be.*[7]
>
> **Dave Chismon**
> Chief technology officer for architecture, UK National Cyber Security Centre

So the real question isn't "should I install OpenClaw." Your fancy Claude assistant (or any AI agent you use) may be sending your documents to malicious actors right now.

---

REFERENCES


1Rehberger, J., *Claude Pirate: Abusing Anthropic's File API For Data Exfiltration*, Embrace the Red, 28 October 2025 - the disclosure, and the "safety protects you from accidents" line - https://embracethered.com/blog/posts/2025/claude-abusing-network-access-and-anthropic-api-for-data-exfiltration/↩

2MITRE ATLAS, *Exfiltration via AI Inference API* - https://atlas.mitre.org/↩

3Lampson, B. W., *A Note on the Confinement Problem*, Communications of the ACM 16:10, 613-615 (1973) - where the class is named - https://doi.org/10.1145/362375.362389↩

4OWASP GenAI Security Project, *LLM01: Prompt Injection* - https://genai.owasp.org/llmrisk/llm01-prompt-injection/↩

5Willison, S., *Prompt injection attacks against GPT-3*, 12 September 2022 - where the attack is named - https://simonwillison.net/2022/Sep/12/prompt-injection/↩

6OpenAI, *Continuously hardening ChatGPT Atlas against prompt injection attacks*, 22 December 2025 - source of "unlikely to ever be fully solved" - https://openai.com/index/hardening-atlas-against-prompt-injection/↩

7Chismon, D., *Prompt injection is not SQL injection*, UK National Cyber Security Centre, 8 December 2025 - https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection↩

CITE THIS ARTICLE

Gurov, P. (2026). *How Claude is hacked through infected Microsoft Word documents*. Gurov Digital Research. https://gurovdigital.com/research/claude-word-document-exfiltration

Published at https://gurovdigital.com/research/claude-word-document-exfiltration · Author ORCID 0009-0001-0152-8056