AI Agent Security Audit Sample Report

Prepared for: Example Co., a 22-person regional services company

Prepared by: Mark Yegge — The AI Blind Spot

Date: November 2026

THIS DOCUMENT IS AN EXAMPLE ONLY.

Example Co. is a fictional company. No client results are shown.

In this report

1. Executive summary
2. Risk scorecard
3. Example findings
4. 30-day fix plan
5. Sources and method

By Mark Yegge

Executive Summary

Example Co. is a 22-person regional services company that uses AI tools in sales, support, marketing and document handling. This audit reviewed the AI tools, agents and third-party services currently in use across the business, and the permissions they hold.

The review found three areas that need attention. None of them required changing any system during the audit: everything below is a finding and a recommended fix, and your team decides what to act on.

What it means in plain English

- One AI assistant is reading more data than the team needs it to, including files outside the folder it works in.
- Two tools can accept instructions from outside content (emails, web pages and pasted text) with no guard against hidden instructions.
- A third-party plug-in holds permissions that have not been reviewed since it was installed.

None of these findings showed any evidence of a breach. They are the kind of exposure that is common in small and mid-size businesses that have grown their AI use quickly — and they are all fixable, usually in a few hours of work.

Headline numbers

- 3 findings rated medium risk or higher.
- 2 of 3 findings are fixable inside a single working day.
- 1 finding needs a vendor or plug-in owner decision.

Statistics in this report are drawn from public third-party sources; see Sources and method. The company, findings and risk scores in this sample are fictional and shown only to illustrate the format of the real deliverable.

By Mark Yegge

Risk Scorecard

Each finding is rated by how likely it is to cause harm and how much harm it could do. Red is fix first.

Finding 1 — Over-broad da	RED	High likelihood · High impact
Finding 2 — Hidden-instruc	AMBER	Medium likelihood · High impact
Finding 3 — Unreviewed pl	AMBER	Medium likelihood · Medium impact

Overall position: AMBER — no breach evidence, three exposures to close.

How this is decided

Likelihood and impact are scored against the way each tool is actually used by your team, not against the maximum possible risk. A to

SAMPLE REPORT

By Mark Yegge

Example Findings

Three findings are shown here to illustrate how the full report presents each issue. A real audit report follows this exact format for every finding in scope.

Finding 1 — Over-broad data access (RED)

Plain English risk: The sales AI assistant is connected to a shared folder that includes past client files, contracts and internal pricing. The team only needs it to read the current sales pipeline.

Evidence: Review of the assistant's configured connections and folder permissions showed read access to 14 folders; 11 are outside its working scope. Permission list screenshot attached.

Fix: Re-scope the assistant to the 3 folders it uses, and add a rule that new folders must be granted explicitly. Owner: internal IT lead. Effort: about 1 hour.

Finding 2 — Hidden-instruction exposure (AMBER)

Plain English risk: Two tools process text from emails and web pages directly. Outside content can contain instructions the tools were not asked to follow.

Evidence: In safe, permission-based tests, both tools followed a benign instruction embedded in sample content. Neither test touched production data or sent anything.

Fix: Turn on the vendor's content-filtering option in both tools, and set a rule that pasted external text is treated as data. Owner: tool owners. Effort: about 2 hours.

Finding 3 — Unreviewed plug-in permissions (AMBER)

Plain English risk: A plug-in installed six months ago holds permission to read and write documents in the company workspace. No one has reviewed what it needs.

Evidence: Plug-in permission inventory shows read/write scope granted at install; usage logs show read-only use.

Fix: Reduce the plug-in to read-only, or re-install with the minimum scope. Owner: vendor account admin. Effort: about 30 minutes.

EXAMPLE REPORT — fictional company. Findings shown are illustrative, not real client results.

By Mark Yegge

30-Day Fix Plan

A real report ends with a plain-English plan. Every fix names who should do it, what ready looks like, and a review date. Nothing in this plan happens automatically — you decide what to act on.

Days 1-7

- Re-scope the sales assistant to the folders it uses (Finding 1).
- Turn on content filtering in the two tools that accept outside text (Finding 2).

Days 8-21

- Reduce the plug-in to read-only or re-install it with minimum scope (Finding 3).
- Write the one-page permission rule: who can connect a new tool, and what it may read.

Days 22-30

- Team review of the rule and a quick re-check of the tools changed this month.
- Decide whether Safe-AI Monitoring should re-check these tools monthly ($297/mo, cancel anytime).

Sources and method

The Business Impact: To show why prompt-injection and over-permissioned AI tools matter, the full report cites public research, including the IBM Cost of a Data Breach Report 2024, which puts the global average cost of a data breach at USD 4.88 million — up 10% from 2023 (IBM, July 2024; ibm.com/reports/data-breach). No statistic in a real report is used without a working link to its source.

Method: inventory of tools and permissions, safe permission-based tests, review of third-party services. The audit is a review and a fix list. It does not include doing the fixes, and no review can find every possible risk.

EXAMPLE REPORT — fictional company. Sources cited are real; client content is not.