

Why AI Changes Your OpSec Surface

For everyone · ~8 min read

The shift from local tools to remote disclosure

For decades, the tools you used at work ran on the machine in front of you. A spreadsheet opened a file on your disk. A text editor saved bytes to a folder. A calculator did math in RAM and forgot it. When you closed the lid, the data stayed where it was. If you wanted to send something to someone else, you had to take an explicit action: attach a file, paste into an email, upload to a share. The default was local, and the default was private.

AI tools have inverted that default. When you type into a chat box powered by a large language model, the words you type do not stay on your device. They are packaged into a network request and sent to a company's servers, where they are processed, logged, and sometimes reviewed by humans or used to improve the model. The interaction looks like a chat window, but the plumbing underneath is a continuous stream of disclosures to a third party. Every prompt is an outbound transmission. There is no "local mode" for a hosted AI service — the local-feeling UI is the whole trick.

What a provider actually sees

When you send a prompt, the provider receives much more than the sentence you typed. It receives a request payload, a bundle of metadata, and a place in a logging pipeline that may route the data through several internal systems. Most users never think about this bundle because the interface only shows them the reply. But the bundle is what determines your exposure.

Here is a realistic inventory of what reaches the provider on a single prompt:

- **The request payload** — the full text of your prompt, any attached files, any prior conversation turns included as context, and any system prompt your employer or tool has configured.

- **Account and identity metadata** — the account or API key that made the call, and through it, your name, email, organization, and billing details.
- **Network metadata** — the IP address of the client, timestamps accurate to the millisecond, User-Agent strings, and sometimes geolocation inferred from IP.
- **Model and routing metadata** — which model you selected, which region handled the call, which feature flags were on, how long the call took.
- **Downstream logging** — the prompt and response recorded in the provider's storage for abuse detection, debugging, quality monitoring, and support ticket investigation.
- **Abuse review** — for prompts that trip safety classifiers, a human reviewer may read the content.
- **Sub-processors** — cloud hosting providers, content moderation vendors, observability platforms, and analytics tools that the AI provider itself relies on. Each is another set of servers and another set of employees.

Notice that none of this is malicious. This is the normal operation of a hosted service. The point is not that providers are lying — it is that the surface area is much larger than a chat window suggests.

Retention, training, and opt-outs

Different tiers of the same product treat your data very differently. On consumer and free tiers, prompts are typically retained by default and may be used to improve the model unless you explicitly opt out in settings. On enterprise or API tiers with a negotiated contract, you may be on a "zero-retention" arrangement where the provider commits not to store prompts beyond the immediate request, and not to use them for training.

The critical word in that sentence is "commits." A zero-retention posture is a contractual promise between your organization and the provider. It is not how the service works by default, and it is not a technical guarantee you can verify from the outside. If you are a staff user clicking around in a web UI, you are almost certainly on a default-retain tier unless your IT or security team has told you otherwise in writing. Training opt-outs are the same story: they exist, they work, but only if someone turned them on and confirmed them. Treat every opt-out as configuration until proven otherwise.

Retention periods also matter. "We delete after 30 days" still means 30 days of a copy sitting in a datastore that is a legitimate target for attackers, insiders, and legal process. Shorter is better. Zero is better still. But no tier, no matter how premium, gives you the same posture as a tool that never sent the data at all.

Logs are leaks

There is a common misreading of AI privacy debates that goes: "The provider promised not to train on my data, so I'm fine." Training is only one of several ways data can escape. Logs held for abuse detection are still logs. Logs held for debugging are still logs. Logs held for "quality monitoring" are still logs. Every retained copy is a potential breach disclosure, a potential subpoena target, and a potential insider threat vector.

Think about it the way a security team does. If an attacker compromises the provider, they do not care whether your prompts were earmarked for training or only for debugging — they dump whatever they can reach. If a court issues a subpoena, it asks for what exists, not for what was tagged as useful. The relevant question is not "will this be used?" but "will this exist?" If the answer is yes, assume it can leak.

The "a prompt is a disclosure" rule

All of the above collapses into a single mental model you can carry into every interaction with an AI tool. Do not think of a prompt as a thought, a search query, or a note to yourself. Think of it as a document you are voluntarily handing to an outside party. Every time you press enter, you are putting a piece of paper in an envelope, addressing it to a company, and dropping it in the mail. You cannot get it back. You do not control who reads it.

The rule

A prompt is a disclosure. If you would not email it to a vendor, do not paste it into an AI.

This rule is deliberately blunt. It is meant to be uncomfortable because the productivity gains of AI tools make it very easy to forget. You will catch yourself about to paste a log file, a customer ticket, a piece of code, or a draft memo. Pause. Ask the question. If the answer is "I'd never forward this to a random vendor by email," the answer to the AI prompt is the same.

Examples in the wild

In early 2023, Samsung engineers pasted internal source code and meeting notes into ChatGPT to help debug and summarize their work, and within weeks Samsung banned the internal use of generative AI tools across the company. The lesson is not that the engineers were reckless or that the tool was malicious; the lesson is that well-intentioned, productive employees using a chat window can move confidential data outside the company in minutes, and once it is gone there is no recall.

Rules to remember

- Treat every prompt as a document you are handing to an outside party.
- Assume prompts are retained, reviewed, and possibly used for training unless you have verified otherwise in writing.
- Metadata travels with the prompt: account, IP, timestamps, project name — not just the text.
- A "zero-retention" claim is a contract, not a default. Check the tier.
- If you would not email it to a vendor, do not paste it into an AI.