

Lesson 04 — General OpSec Rules for AI

The master cheat-sheet: ten everyday rules, one response protocol, and the line between a near-miss and an incident.

Audience: Everyone

AI OpSec

Cheat Sheet

Notes: Summary lesson. Pin the rules somewhere visible and come back to them whenever you open an AI tool.

What you will learn

- **The cheat sheet** — ten rules you can recite from memory
 - **When to escalate** — the difference between a near-miss and a reportable incident
 - **Stop-drop-tell** — the three-step response when you realise you pasted something you shouldn't have
-

The ten everyday rules

1. Classify every prompt before you send it.
 2. Never mix personal and corporate AI accounts or identities.
 3. Approved tools only for corporate work. Personal tools only for personal work.
 4. Redact values, keep shapes.
 5. Assume prompts are retained, reviewed, and breachable.
 6. Treat screenshots, file uploads, and IDE autocomplete as prompts.
 7. One browser profile per jurisdiction. No cross-login.
 8. Stop-drop-tell on a mistake. Do not delete the chat.
 9. Rotate any secret you suspect touched an AI — even "maybe".
 10. When in doubt, ask the security team before pasting.
-

Approved vs. unapproved tools

"Approved" is a legal status, not a preference

A tool is approved when your employer has a contract with the provider, a zero-retention or DPA clause, and an audit trail the admins can review.

- **Contract** — an enterprise agreement, not a consumer ToS
 - **Zero-retention** — prompts and responses are not stored beyond the request
 - **Admin audit** — your security team can see who used what, when
 - **Acceptable-use policy** — written rules you agreed to at onboarding
-

Near-miss vs. incident

Near-miss

You almost pasted something sensitive, or you did and caught it within minutes and reported it before any harm. No credential was used, no data moved downstream.

Incident

A credential, PII, customer data, or production secret reached a tool — and either stayed there, was missed, or was discovered later. Must be reported and investigated.

When in doubt, report it as an incident. Your security team would rather triage a near-miss than discover a real leak a week later.

Stop-drop-tell

Three steps, in order

1. **Stop** using the tool. Do not send another prompt in that session.
2. **Capture** the evidence. Screenshot the prompt and the response.
3. **Tell** security. Use the reporting channel your company has set up.

Do not delete the chat. The provider still has it, and deletion destroys the evidence your security team needs to triage.

What NOT to do after a mistake

- **Don't delete the chat** — it stays in provider logs anyway, and you lose the evidence
 - **Don't try to "retract"** — there is no recall button; the provider already has the bytes
 - **Don't wait** — minutes matter; a fast report turns an incident into a near-miss
 - **Don't hide it** — security teams reward reporters; they do not punish them
-

Example: near-miss

Fast report, clean outcome

A developer pastes an internal staging URL into ChatGPT to ask about a 500 error. They realise within thirty seconds, stop, screenshot the chat, and message the security channel.

- Reported in under 5 minutes
 - No credential, no customer data, no secret
 - Security logs it as a near-miss and moves on
 - Dev gets a "thanks for reporting" — not a write-up
-

Example: escalation

Silent mistake, real incident

A developer pastes a JWT into a consumer AI tool to "decode" it. They say nothing. A week later the same token is used from an unfamiliar IP against the production API.

- Rotate the token and every related credential
 - Full investigation: logs, sessions, downstream access
 - Disciplinary review — not for the mistake, for the silence
 - The cost of reporting was zero. The cost of hiding it was everything.
-

Rules to remember

1. Classify every prompt before you send it.
 2. Never mix personal and corporate AI accounts or identities.
 3. Approved tools only for corporate work. Personal tools only for personal work.
 4. Redact values, keep shapes.
 5. Assume prompts are retained, reviewed, and breachable.
 6. Treat screenshots, file uploads, and IDE autocomplete as prompts.
 7. One browser profile per jurisdiction. No cross-login.
 8. Stop-drop-tell on a mistake. Do not delete the chat.
 9. Rotate any secret you suspect touched an AI — even "maybe".
 10. When in doubt, ask the security team before pasting.
-

