Using ChatGPT and other generative AIs safely

Cloud-based AI services—like ChatGPT, Google Gemini, Claude, and Perplexity—
carry the usual risks associated with the cloud, and some unique generative-AI hazards.

Input Risks

Anything you share with a cloud-based AI leaves your network.

Generative AI can process text, speech, images, and documents—and if these contain personal data, proprietary information, or intellectual property, they may invoke cross-border transfer rules, corporate policies, or laws like HIPAA, GLBA, CCPA, and GDPR.

Anything you share can be used to train the AI; training data can reappear in its outputs or be deliberately extracted.

How to protect against input risks:

- ✓ Block unapproved AI tools with DNS filtering.
- ✓ Understand where, why and how long AI tools store data.
- ✓ Choose plans that do not use your data for training.
- ✓ Enforce a generative AI acceptable use policy.
- ✓ Include AI in your security awareness training.
- ✓ Provide prompt templates and cheat sheets.
- ✓ Invoke CCPA/GDPR rights to delete leaked data.

In 2024, researchers successfully extracted 2,702 credentials accidentally shared with GitHub Copilot.[1]

Output Risks

AI relies on learned patterns, not facts, so its outputs can be biased, misleading, inconsistent, or include plausible but false "hallucinations."

It also shows quirks—overusing words like "realm" or "elevate" in text, or drawing hands with too many fingers in pictures.

In the US, purely AI-generated works aren't entitled to copyright.

How to protect against output risks:

- ✓ Include AI in your security awareness training.
- ✓ Ask AI tools to provide sources and verify them manually.
- ✓ Mandate two reviewers for customer content.
- ✓ Audit AI-generated code for common insecurities and errors.
- ✓ Get sign-off for any AI output used in critical business decisions.

In 2023, SDNY imposed a $5,000 fine on an attorney for using fake judicial opinions hallucinated by ChatGPT.[2]

Protect your business against AI-powered cyberattacks. Visit **threatdown.com**

[1] Yizhan Huang et al (2024), Your Code Secret Belongs to Me: Neural Code Completion Tools Can Memorize Hard-Coded Credentials, https://arxiv.org/pdf/2309.07639
[2] Justia (2023), Mata v. Avianca, Inc., No. 1:2022cv01461 - Document 54 (S.D.N.Y. 2023), https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/

THREATDOWN™