AI Hallucinations & Data Privacy

A Quick Guide for AI Users

What hallucinations are, how to spot and reduce them, and how to protect your personal and company data across popular AI tools.

Part 1: Understanding AI Hallucinations

What Is an AI Hallucination?

An AI hallucination occurs when a generative AI tool produces information that sounds plausible and confident but is factually incorrect, fabricated, or misleading. The AI is not deliberately lying, it is generating the statistically most likely next words based on patterns in its training data. When it encounters gaps in its knowledge or ambiguous prompts, it fills those gaps with plausible-sounding content rather than saying “I don’t know.”

Think of it like a very confident colleague who always gives you an answer, even when they’re not sure, they’ll piece together something that sounds right rather than admit uncertainty.

Common Types of Hallucination

- **Fabricated facts:** The AI invents statistics, dates, quotes, or events that never happened. For example, citing a research paper that does not exist, or attributing a quote to someone who never said it.
- **False citations:** Generating references, URLs, or source attributions that look authentic but link to non-existent pages or papers.
- **Confident nonsense:** Providing detailed, authoritative-sounding answers to questions it cannot possibly know the answer to.
- **Outdated information:** Presenting information that was correct at the time of training but is no longer accurate, without flagging that it may be outdated.
- **Blended facts:** Combining real information from different contexts to create a plausible but inaccurate composite answer.

Why Does It Happen?

AI language models work by predicting the most probable next word in a sequence. They do not “understand” content the way humans do, they identify patterns. Hallucinations occur because the models are trained to always produce an answer (they are rewarded for fluent output, not for admitting uncertainty), training data may contain gaps, biases, or errors, complex or ambiguous prompts can push the model outside its reliable knowledge, and the model has no real-time connection to facts unless it has web search enabled.

Research suggests that AI chatbots hallucinate between roughly 2.5% and 22% of the time, depending on the model, the task, and how the question is framed.

> ⚠ **Real-World Example**
>
> In 2023, a US lawyer used ChatGPT to prepare a legal brief. The AI generated several case citations that looked completely authentic but were entirely fabricated, the cases did not exist. The lawyer submitted the brief without checking, resulting in sanctions from the court. This case became a widely-cited cautionary tale about the dangers of trusting AI output without verification.

How to Reduce the Risk of Hallucinations

You cannot eliminate hallucinations entirely, but you can significantly reduce them with the right habits:

1. **Be specific in your prompts.** Vague questions lead to vague (and often invented) answers. Provide context, constraints, and clear instructions. For example, instead of "Tell me about UK employment law," try "Summarise the key changes to UK unfair dismissal rules introduced in 2024, citing specific legislation."
2. **Ask the AI to show its reasoning.** Prompting the AI to explain its thinking step-by-step reduces logic leaps that lead to hallucinations. Add "Explain your reasoning" or "Show your working" to your prompt.
3. **Tell it to say "I don't know."** Include instructions like "If you are not confident in the answer, say so rather than guessing." This gives the AI permission to be uncertain rather than fabricating.
4. **Use web search features.** Tools like Perplexity, ChatGPT with browsing, Gemini, and Claude all have web search capabilities that ground responses in current, citable information. Always prefer sourced answers over pure generation.
5. **Cross-check and verify.** Never publish or act on AI-generated facts, statistics, legal references, or medical information without independent verification. This is the single most important habit you can build.
6. **Use source-grounded tools for research.** NotebookLM only answers from your uploaded sources. Perplexity cites every claim. These tools are inherently less prone to hallucination than open-ended chatbots because they are anchored to specific information.
7. **Lower the "temperature" if available.** Some tools let you adjust creativity settings. Lower creativity (temperature) means more deterministic, factual output. Higher creativity means more inventive but less reliable answers.
8. **Keep prompts focused and concise.** Constraining the scope and length of requested responses reduces the opportunity for the AI to wander into fabricated territory. Break complex requests into smaller, verifiable steps.

The Golden Rule

Treat AI output like advice from a very knowledgeable but occasionally unreliable colleague. Use it as a starting point, not a final answer. Always verify anything that matters before you share it, publish it, or act on it.

Part 2: Protecting Your Data & Privacy

Why This Matters

By default, most AI tools use your conversations to improve their models. This means anything you type, including company information, client details, personal data, or intellectual property, could become part of the AI's training data. Once information is incorporated into a model's training, it cannot be removed.

This is particularly important for business users. If you or your staff are using free or individual paid plans for work tasks, you may be inadvertently exposing sensitive commercial information unless you explicitly opt out of training.

> ⚠ **Critical Warning for Business Users**
>
> Paying for a "Pro" or "Plus" individual plan does NOT automatically protect your data. On most platforms, individual paid plans still default to training on your data, you are paying for extra features, not for privacy. Only Business/Team/Enterprise plans provide contractual data protection by default. If your staff are using personal AI subscriptions for work, their conversations may be training the AI models.

How to Switch Off Training: The Big Four Chatbots

Privacy Setting	ChatGPT	Google Gemini	Microsoft Copilot	Claude
Training on by default? (Free/Paid Individual)	YES – on by default for Free, Plus and Pro plans	YES – on by default; saved for 18 months	Consumer: YES (with identifiers removed). Enterprise: NO	YES – on by default (changed Sept 2025; was previously off)
How to switch training OFF	Settings → Data Controls → Toggle OFF "Improve the model for everyone"	Go to myactivity.google.com/product/gemini → Turn OFF "Keep Activity" (formerly Gemini Apps Activity)	Consumer: Profile → Privacy → Toggle off training on text and voice. Enterprise: Protected by default	Settings → Privacy → Toggle OFF "Help improve Claude"
Data retention after opt-out	30 days (for safety/abuse monitoring)	Up to 72 hours for abuse detection. Previously saved data may be retained up to 3 years	Consumer: up to 18 months chat history (adjustable). Enterprise: governed by admin policies	30 days retention after opt-out. If opted in, up to 5 years
Business/Team accounts	Team & Enterprise: NOT trained on by default (contractual DPA)	Workspace Business: NOT trained on by default; treated like other Workspace data	Enterprise M365 Copilot: NOT trained on by default; commercial data protection	Team & Enterprise: NOT trained on by default

Incognito / Temp Chat	Temporary Chat available – not saved, not trained on (deleted after 30 days)	Temporary Chat feature – not saved, not used for training	No specific incognito mode for consumer Copilot	Incognito Conversations available – not saved, not trained on
Can you delete chat history?	Yes – Settings → General → Delete all chats	Yes – via Google Account activity controls	Yes – via Copilot activity in Microsoft account	Yes – delete individual chats or all history
Human review of chats?	Possible – for safety monitoring and model improvement	Possible – conversations may be reviewed by human evaluators	Enterprise: No. Consumer: possible for quality/safety	Possible – flagged conversations may be reviewed for safety
Key warning	Giving thumbs up/down feedback may explicitly opt that conversation into training	Turning off activity disables chat history and some extension features	Consumer Copilot has weaker privacy protections than enterprise version	Flagged conversations and explicit feedback are always retained for safety regardless of opt-out

How to Switch Off Training: Specialist Tools

Privacy Setting	Napkin.ai	Perplexity.ai	NotebookLM
Data used for training?	Data shared with AI subprocessors to provide functionality. You can opt out of data sharing; contractual agreements prevent subprocessors training on your data	Free tier: data may be used to improve service. Pro: additional privacy controls. Enterprise: data not used for training	Via Google Workspace: NOT trained on. Personal accounts: governed by Google's Gemini policies (activity settings apply)
How to protect your data	Opt out of data sharing in account settings. Your content remains yours regardless of plan	Review privacy settings in your Perplexity account. Enterprise users get contractual data protection	Use a Google Workspace account for full protection. For personal accounts: turn off Keep Activity at myactivity.google.com
Key consideration	Your uploaded documents (PPT, PDF, Word) are processed by AI to generate visuals – be cautious with confidential materials	Your search queries and uploaded files are processed in the cloud. Free tier has fewer privacy controls	Your uploaded sources are processed by Google's Gemini AI. Workspace accounts have enterprise-grade protection; personal accounts follow consumer Google policies

Practical Data Protection Checklist

9. **Switch off training on every AI tool you use.** Do this today, before your next conversation. Follow the steps in the table above for each platform.
10. **Use Incognito/Temporary Chat for sensitive topics.** Claude, ChatGPT, and Gemini all offer temporary conversation modes that are not saved or trained on. Use these for anything confidential.
11. **Never share truly sensitive information.** Regardless of your settings, do not enter passwords, financial account details, client personal data, or trade secrets into any AI chatbot. Once submitted, you have lost control of that information.
12. **For business use, insist on Team/Enterprise plans.** These plans provide contractual guarantees that your data will not be used for training. Individual "Pro" plans do not offer this protection by default.
13. **Check your settings quarterly.** AI companies rename, restructure, and sometimes reset privacy controls without fanfare. What you set three months ago may not be what's active today.
14. **Be cautious with feedback buttons.** On ChatGPT, clicking thumbs up or thumbs down may explicitly opt that specific conversation into the training pool, even if you have switched off general training.
15. **Anonymise data before uploading.** If you need to analyse documents containing personal data, remove or anonymise identifiable information before uploading to any AI tool.
16. **Create an AI usage policy for your organisation.** Define which tools are approved, what data can be shared, and ensure all staff have training on safe AI use. This is becoming a regulatory expectation under the EU AI Act.

> **Remember**
>
> You cannot remove data from AI models that have already been trained on it, but you can stop companies from using your future conversations. Take five minutes today to check your settings on every AI tool you use. Your future self (and your clients) will thank you.

Disclaimer: *This guide reflects privacy settings and policies as of March 2026. AI providers frequently update their terms, features, and data handling practices. Always verify current settings and policies directly with each provider. This document is for educational purposes and does not constitute legal advice.*