HEARPRENEUR
SOLUTIONS
THRIVE WITH VALUES

AI Hallucinations

What Every Business Owner Needs to Know in 2026

Understanding the risks, the progress, and the practical steps to use AI safely

February 2026

www.hearpreneur.com.au

Prepared by Hearpreneur Solutions
Empowering hearing clinic owners to thrive with values

Contents

This article contains detailed benchmarks, model-by-model comparisons, practical strategies, and over 50 verified source links.

What Are AI Hallucinations?

AI hallucinations occur when artificial intelligence tools like ChatGPT, Claude, Gemini, or Copilot generate information that sounds confident and plausible but is factually incorrect, fabricated, or misleading. The AI isn't "lying" — it predicts the most likely next words based on patterns in its training data. When it encounters gaps in its knowledge, it fills them in with convincing-sounding content rather than saying "I don't know."

Think of it like an enthusiastic new employee who, rather than admitting they're unsure, confidently provides an answer they've pieced together from memory. Most of the time they're helpful and accurate — but sometimes they get it wrong, and they do so with complete confidence.

The Real Cost of Unchecked AI: Cautionary Tales

The $440,000 Government Report Riddled with Fake References (Australia, 2025)

In one of the most significant AI hallucination scandals to date, Deloitte Australia was paid approximately AUD $440,000 to produce a 237-page review of the Targeted Compliance Framework (TCF) for the Department of Employment and Workplace Relations (DEWR). The report, published in July 2025, was found by Sydney University researcher Chris Rudge to be "full of fabricated references."

A DEWR review identified **dozens of potential errors**, including:

- Citations to non-existent academic books and research papers
- A misattributed quote from a Federal Court judgment
- Incorrect legal references and non-existent laws
- A misnamed judge

Deloitte admitted to using Azure OpenAI's GPT-4o model in drafting parts of the report without initially disclosing this to the client. The firm's revised methodology disclosure stated they used *"a generative AI large language model (Azure OpenAI GPT-4o) based tool chain licensed by DEWR and hosted on DEWR's Azure tenancy"* to fill "traceability and documentation gaps." The firm issued a corrected version in October 2025 and refunded just **AUD $97,587** — less than a quarter of the $440,000 contract. The partner responsible for the report left the firm in December 2025.

Sources: Fortune; The Register; CFO Dive

Lawyers Submitting Fake Court Cases

In 2023, New York lawyer Steven Schwartz used ChatGPT to draft a legal brief containing **six completely fabricated court cases** (including the now-famous "Varghese v. China Southern Airlines"). He and his firm were sanctioned $5,000. By late 2025, **hundreds of documented cases globally** (well over 500 and still climbing, according to Damien Charlotin's AI Hallucination Cases database) of lawyers submitting AI-generated filings with misrepresented or invented case law had been recorded.

A Stanford University study found that when asked verifiable questions about federal court cases, **ChatGPT-4 hallucinated 58% of the time and Llama 2 hallucinated 88% of the time**.

Source: Stanford RegLab & HAI, *Journal of Legal Analysis*, 2024. Hallucinating Law (Stanford)

AI Academic Papers With Fake Citations (January 2026)

In a deeply ironic twist, GPTZero analysed 4,841 papers accepted by NeurIPS 2025 — one of the world's most prestigious AI conferences — and found **100+ confirmed hallucinated citations across 51 papers**. These included author names like "John Doe and Jane Smith," real author names paired with fake paper titles, and fabricated DOIs.

Source: GPTZero NeurIPS Analysis; Fortune; TechCrunch

Other Notable Incidents

Year	Incident	Impact
2024	**Air Canada chatbot** gave incorrect bereavement fare advice	Tribunal ordered $812 compensation — set legal precedent for chatbot liability
2023	**Chevy dealership chatbot** "agreed" to sell a Tahoe for $1	Dealership shut down the chatbot; exposed AI sales vulnerabilities
2025	**Chicago Sun-Times** published AI-generated book list	Fabricated book titles and summaries; led to retractions
2024-25	**Pieces Technologies** (healthcare AI) claimed <0.001% hallucination rate	Found to be false and misleading; company required to disclose limitations
2025	**CEO used AI to summarise investment documents**	Fabricated a non-existent term; nearly derailed a multimillion-dollar deal

How Bad Is the Problem? Current Benchmarks

The hallucination picture in 2026 has two very different sides. When AI is given source material to work with (like summarising a document you provide), top models are impressively accurate. But when asked to recall facts or reason about complex topics from memory alone, error rates climb significantly.

Table 1: Grounded Summarisation Tasks — Hallucination Rates (Vectara Leaderboard, Feb 2026)

These rates measure how often AI adds unsupported information when summarising a provided document — the safest use case.

Model	Provider	Hallucination Rate	Accuracy
Gemini 2.0 Flash	Google	0.7%	99.3%
Gemini 2.0 Pro	Google	0.8%	99.2%
o3-mini-high	OpenAI	0.8%	99.2%
GPT-4o	OpenAI	1.5%	98.5%
finix_s1_32b	AntGroup	1.8%	98.2%
Gemini 2.5 Flash Lite	Google	3.3%	96.7%
Phi-4	Microsoft	3.7%	96.3%
Llama 3.3 70B	Meta	4.1%	95.9%
Claude Sonnet	Anthropic	4.4%	95.6%
Claude Opus	Anthropic	10.1%	89.9%

Source: Vectara Hallucination Leaderboard, updated Feb 2026. Tests 98 models using HHEM-2.3 evaluation.

> ***What this means:*** *When you paste a document into AI and ask it to summarise, the best models get it right over 99% of the time. This is where AI shines.*

Table 2: Open-Domain & Complex Tasks — Hallucination Rates (Various Benchmarks, 2025-2026)

These rates measure performance on harder tasks: answering questions from memory, reasoning, searching the web, or working in specialised domains.

Task Type	Typical Hallucination Range	Context
General factual recall	8–35%	Varies significantly by model and topic
Legal questions	58–88%	Stanford study, 2024
Medical/clinical prompts	44–82%	Adversarial and decision-support scenarios
Web search accuracy	6–76%	Depends heavily on the tool used
News citation accuracy	6–94%	Columbia Journalism Review, March 2025
Multilingual translation	33–60%	Even top models struggle

Task Type	Typical Hallucination Range	Context
Complex reasoning/analysis	15–43%	Statement analysis tasks

Sources: Stanford RegLab; Columbia Journalism Review, March 2025; AIMultiple; Relum workplace study, Dec 2025.

> ***What this means:*** *When AI has to "remember" facts or reason through complex problems without source documents, error rates can be substantial. This is where verification matters most.*

Table 2b: Open-Domain Factual Recall — SimpleQA (OpenAI, 2024–2025)

SimpleQA is OpenAI's own factuality benchmark for short, verifiable factual questions where there is a single correct answer. Each response is graded as correct, hallucinated (incorrect but attempted), or not attempted. These figures come from OpenAI's SimpleQA paper and release materials.

Model	Provider	Correct Answers	Hallucination Rate	Notes
GPT-4.5	OpenAI	~62–63%	**37.1%**	Newer flagship; substantially better than GPT-4o on SimpleQA
GPT-4o	OpenAI	~38%	**61.8%**	OpenAI's chart shows it hallucinating on roughly 3 out of 5 SimpleQA questions
o1	OpenAI	~56% (est.)	~44% (est.)	First reasoning model; hallucination rate around mid-40s in early benchmarking
o3-mini	OpenAI	~20%	**80.3%**	Smaller reasoning model; hallucinating on around 4 out of 5 SimpleQA questions

Sources: OpenAI — Introducing SimpleQA; OpenAI — SimpleQA Paper (PDF); Simon Willison — Introducing GPT-4.5; Helicone — GPT-4.5 Benchmarks

> ***The contrast is stark:*** *The same GPT-4o that shows around **1.5%** hallucination on tightly constrained, document-grounded summarisation tasks (Vectara leaderboard, Table 1) has a **61.8% hallucination rate** on SimpleQA when asked to answer short factual questions from memory. When GPT-4o is summarising a document you give it, almost everything it says can be traced back to the source text. But when asked to recall facts without external documents, it makes things up more often than it gets them right. GPT-4.5 improves this considerably (down to 37.1%), but still hallucinates more than one answer in three.*
>
> ***What this means for you:*** *"Ask it from memory with no supporting documents" is the riskiest way to use AI. Whenever the stakes are non-trivial, you're far better off either giving the model real documents to work from (uploads, pasted text, or RAG-enabled tools like Perplexity) or using AI as a drafting partner whose outputs you then verify — rather than as an oracle expected to "just know" the answer.*

Table 3: Galileo Hallucination Index — Overall Accuracy Rankings (2024)

The Galileo index evaluates models across Q&A, RAG-assisted Q&A, and long-form generation tasks.

Ranking	Model	Key Finding
#1 Most Accurate	Claude 3.5 Sonnet (Anthropic)	Highest factual consistency across all task types
Best Value	Gemini 1.5 Flash (Google)	~10x cheaper than Claude 3.5, scored 0.92–1.0 accuracy
22 models evaluated	Including GPT-4, Llama, Mistral variants	All showed measurable hallucination in at least one task

Source: Galileo Hallucination Index, 2024 edition. Evaluates using ChainPoll methodology across 7 benchmark datasets.

Not All Benchmarks Are Created Equal: The Claude Paradox

If you've looked at the tables above, you may have noticed something puzzling. On the Vectara leaderboard (Table 1), Claude models appear to be among the *worst* performers. But on the Galileo index (Table 3), Claude 3.5 Sonnet was ranked the **#1 most accurate LLM in the world**. On the Artificial Analysis Omniscience Index, Claude Opus 4.6 — the latest and most powerful Claude model, released February 2026 — ranked **2nd globally** for factual knowledge.

How can the same model family be both the best and worst at hallucinating?

The answer reveals something every AI user should understand: **what a benchmark measures matters as much as the number it produces.**

The Same Model, Wildly Different Scores

Benchmark	What It Measures	Claude Result	Rank
Galileo Index 2024	Accuracy in RAG-assisted Q&A	0.97–1.0 scores	**#1 of 22 models**
AA-Omniscience	Factual knowledge + knowing when to refuse	Score: 11	**#2 globally**
TruthfulQA	Resistance to common misconceptions	~2x improvement over Claude 2.1	Top tier
Vectara Leaderboard	Strict document summarisation	4.4–12.2% hallucination	Bottom half

Sources: Galileo Index; Artificial Analysis — Opus 4.6; Vectara Leaderboard

What's Actually Happening

The Vectara benchmark gives AI a document and says: *"Summarise using ONLY the information in this passage. Do not infer. Do not use your internal knowledge."*

Claude is trained to be *helpful*. When it summarises a document, it instinctively enriches the summary with relevant context from its own vast knowledge — adding true, useful information that wasn't in the source text. A study (FaithBench, NAACL 2025) found that **21.3% of Claude's summaries contain "benign hallucinations"** — information that is factually correct but not present in the source document.

On Vectara's benchmark, adding *true but not-in-the-source* information still counts as hallucination. So Claude gets penalised for being too helpful — for adding context that a human reader might actually appreciate.

Meanwhile, the Galileo benchmark tests whether Claude sticks to retrieved context when answering questions — a scenario where Claude's strong instruction-following shines. And the AA-Omniscience index rewards models that **know when to say "I don't know"** — something Claude does more than almost any other model.

Source: FaithBench (arXiv); Benchmarking LLM Faithfulness in RAG (arXiv)

Claude Opus 4.6: The Latest Data

The newest Claude model (Opus 4.6, released February 5, 2026) shows this same pattern:

Vectara Leaderboard — Claude Model Family (Grounded Summarisation):

Model	Hallucination Rate
Claude Haiku 4.5	9.8%
Claude Sonnet 4	10.3%
Claude Opus 4.1	11.8%

Model	Hallucination Rate
Claude Sonnet 4.5	12.0%
Claude Opus 4.6	12.2%

AA-Omniscience Index — Overall Factual Knowledge:

Model	Provider	Score	Rank
Gemini 3 Pro Preview	Google	13	1st
Claude Opus 4.6	**Anthropic**	**11**	**2nd**
Claude Opus 4.5	Anthropic	10	3rd

The pattern is consistent: Claude Opus 4.6 is the 2nd most knowledgeable and factually accurate model in the world, but it ranks poorly on strict summarisation because it adds helpful context beyond the source material.

Source: Artificial Analysis — AA-Omniscience; Vectara Leaderboard (GitHub)

The Refusal Factor

There's another important nuance. On Google's FACTS factuality benchmark (which tests closed-book factual recall), Claude 4 Sonnet refused to answer about **45% of parametric knowledge questions** rather than guessing — far more than competitors like GPT or Gemini. This is by design: Anthropic's CEO Dario Amodei stated in May 2025 that *"AI models probably hallucinate less than humans, but they hallucinate in more surprising ways."*

This doesn't mean Claude refuses 45% of all everyday prompts — it's specifically when asked to recall facts from memory that it is most likely to say "I'm not sure" rather than guess. When Claude *does* answer (rather than refusing), it tends to give detailed, enriched responses — which means its per-response hallucination rate on Vectara is higher. Models like Gemini that attempt everything with minimal refusal generate tighter, more constrained summaries and score lower on Vectara's metric.

But from a user's perspective, would you rather have a model that confidently gives you wrong information, or one that tells you when it's not sure?

Sources: Anthropic CEO on Hallucinations (TechCrunch); Why Claude Refuses 45% of Questions (B2B News Network)

The Benchmark Scoring Model Has Limitations Too

It's worth noting that the scoring model itself (Vectara's HHEM) isn't perfect. A 2025 arXiv study found that HHEM produced **16 ranking inversions** compared to human judgements — meaning it ranked models in the wrong order 16 times. Even the best hallucination detection models achieve only about **50% accuracy** on genuinely challenging cases.

Different evaluation methods can produce **contradictory conclusions** about the same model. For example, HHEM rates GPT-o1 as having a higher hallucination rate than GPT-4o, while another evaluation method (FACTS) shows the exact opposite.

Source: Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards (arXiv)

What This Means for You

This section isn't about declaring one model "better" than another. It's about a critical lesson for anyone reading AI benchmarks:

1. **Always ask "what exactly is being measured?"** A model with a 12% hallucination rate on one benchmark can be the #1 most accurate on another — because they test different capabilities.
2. **No single number tells the whole story.** A model that refuses uncertain answers may look worse on some benchmarks but is actually safer to use.
3. **Match your tool to your task.** If you need strict document summarisation with zero additions, prompt explicitly for that. If you need accurate knowledge work and honest uncertainty, different strengths matter.
4. **Be sceptical of headlines.** "Model X hallucinates 12 times more than Model Y" may be technically true on one narrow benchmark while being misleading about real-world reliability.

The Good News: Significant Progress Over Time

Table 4: How Hallucination Rates Have Improved (2023–2026)

Period	Best Models (Grounded Tasks)	Best Models (Open-Domain)	Key Milestone
Early 2023	~8–15%	30–50%+	"Hallucinate" becomes Dictionary.com Word of the Year
Late 2023	~5–10%	25–40%	First Galileo Hallucination Index published
Mid 2024	~2–5%	15–35%	RAG becomes standard enterprise approach
Early 2025	~1–3%	10–25%	Vectara leaderboard expands to 98 models
Early 2026	~0.7–2%	8–20%	Sub-1% achieved for grounded summarisation

> ***The trend is clear:*** *For controlled tasks with source material, hallucination rates have dropped from ~15% to under 1% in just three years. Open-domain tasks still lag behind but are improving steadily.*

Why Do AI Models Hallucinate?

Understanding *why* helps you know *when* to be extra careful:

1. **Pattern completion, not understanding** — AI predicts probable next words. It doesn't truly "know" facts; it recognises patterns. When patterns are ambiguous, it guesses.
2. **Trained to be confident** — Models are rewarded for providing complete, confident answers rather than saying "I'm not sure." This means they rarely refuse to answer even when they should.
3. **Knowledge cutoff dates** — Models are trained on data up to a certain point. Anything after that date is unknown territory, increasing hallucination risk.
4. **Rare or specialised topics** — The less common the topic in training data, the higher the hallucination risk. Obscure legal cases, niche medical conditions, and local business information are particularly vulnerable.
5. **Long or complex requests** — The more steps in a task, the more opportunities for errors to compound.

The Confidence Paradox

A study of AI medical chatbots found they used substantially more "strong confidence" language — words like "definitely," "certainly," and "without doubt" — when generating *incorrect* information compared to when providing factual answers. One analysis reported this confidence language was **around a third more frequent** in incorrect responses. This inverse relationship between confidence and accuracy makes hallucinations particularly dangerous: the more wrong the AI is, the more convincing it sounds.

Sources: Reported via secondary analyses of academic work linked to MIT, January 2025. See also arXiv — Overconfident LLMs for related cross-lingual overconfidence findings.

The Reasoning Model Paradox

One of the most surprising findings of 2025 was that **more powerful "reasoning" AI models actually hallucinate more, not less**. OpenAI's newer reasoning models showed dramatically higher hallucination rates on factual recall tasks:

Model	Type	Hallucination Rate (PersonQA)
o3-mini	Reasoning	14.8%

Model	Type	Hallucination Rate (PersonQA)
o1	Reasoning	16.0%
o3	Advanced Reasoning	33.0%
o4-mini	Advanced Reasoning	48.0%

The hypothesis is that extended reasoning gives models more opportunities to generate convincing-but-incorrect chains of logic. OpenAI acknowledged in their technical report that "more research is needed" to understand why.

In September 2025, OpenAI published their own research paper ("Why Language Models Hallucinate") arguing that hallucinations persist because **evaluation benchmarks reward guessing over acknowledging uncertainty** — a structural problem in how the entire industry builds and measures AI.

Sources: TechCrunch — OpenAI's Reasoning Models Hallucinate More; OpenAI — Why Language Models Hallucinate

How to Minimise AI Hallucinations: Practical Strategies

For Everyday Users

Strategy	How It Works	Effectiveness
Use the RACE Framework	Give AI a Role, Action, Context, and Expectations — structured prompts significantly reduce errors	Studies show structured prompting with clear roles and constraints meaningfully reduces hallucinations
Ask for sources	Add "cite your sources" or "provide references" to your prompts	Forces the AI to ground answers; makes fabrications obvious
Provide context	Paste in your own documents, data, or facts for the AI to work with	Drops rates from 15–35% to 0.7–5% (grounded vs ungrounded)
Add "If unsure, say so"	Include "If you're not completely sure, say 'I'm uncertain'" in your prompt	Significantly reduces confident-sounding errors
Ask "Are you sure?"	In internal experiments, Google researchers found that asking an AI "Are you hallucinating right now?" reduced errors by about 17% in subsequent responses	Simple self-check technique
Use citation-focused tools	Perplexity, Google AI Overview, or ChatGPT with Browse automatically cite sources	Easier to verify; built-in grounding

Strategy	How It Works	Effectiveness
Verify key facts	Spot-check statistics, names, dates, and quotes — especially for anything you'll publish	Takes seconds; catches the most damaging errors
Break complex tasks into steps	Instead of one huge request, use multiple focused prompts	Each step has less room for compounding errors

For Business Owners

Strategy	Description	Effort Level
Create AI usage policies	Define which tasks are appropriate for AI and which require human review	Low — one-time setup
Never publish unreviewed AI content	Establish a "human checks before publish" rule for all external-facing content	Low — cultural habit
Use RAG-enabled tools	Tools that search your own documents before generating answers (e.g., Perplexity, enterprise chatbots)	Medium — tool selection
Domain-specific verification	For legal, medical, or financial content, always have a qualified professional review	Low — existing practice
Keep AI for strengths	Drafting, brainstorming, summarising, formatting — tasks where errors are easy to spot and fix	Low — smart task selection

Understanding Retrieval-Augmented Generation (RAG)

RAG is the most effective technical approach for reducing hallucinations. In simple terms:

Without RAG: You ask AI a question → It answers from memory (which may have gaps) → Higher hallucination risk.

With RAG: You ask AI a question → It first searches a knowledge base of real documents → It generates an answer grounded in those documents → Much lower hallucination risk.

How RAG Helps

Aspect	Without RAG	With RAG
Hallucination rate	15–35% (open-domain)	0.7–6% (with quality sources)
Knowledge currency	Limited to training cutoff	Can access current information

Aspect	Without RAG	With RAG
Source verification	No built-in citations	Often provides clickable sources
Example tools	Standard ChatGPT, Claude	Perplexity, ChatGPT Browse, Copilot

Source: Meta AI, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020 (founding paper). Enterprise adoption data from 2025-2026 benchmarks.

Perplexity: The Best AI Search Tool (That's Still Wrong 37% of the Time)

Perplexity deserves special attention because it represents RAG in action — it's what happens when you build an AI tool around searching and citing real sources rather than relying on memory. For small business owners who want an AI research assistant with built-in fact-checking, Perplexity is the most relevant tool to understand.

How Perplexity Differs From ChatGPT, Claude, and Gemini

Feature	Perplexity	ChatGPT / Claude / Gemini
Default behaviour	Always searches the web first	Answers from training data (memory) by default
Citations	Mandatory — every claim gets a numbered source	Optional, often absent or vague
Real-time data	Always current	Depends on whether user enables browsing
Knowledge source	Retrieved web pages	Parametric memory (training data)
Hallucination approach	Grounded in specific documents	Relies on model confidence

This architecture is why Perplexity consistently outperforms other AI tools on factual accuracy in search tasks.

Perplexity's Hallucination Rates

The most rigorous independent test was the **Columbia Journalism Review / Tow Center study (March 2025)**, which ran 200 test queries across 8 AI search engines (1,600 total queries). Each test provided a quote from a published article and asked the AI to identify the title, publication date, publisher, and URL.

Results — Failure Rates by AI Search Tool:

AI Search Tool	Failure Rate
Perplexity (free)	**37%** (best)
ChatGPT Search	~42%
DeepSeek Search	~58%
Gemini	High (created more fabricated links than correct ones)
Copilot	Variable (declined to answer majority of questions)
Grok-2 Search	~82%
Grok-3 Search	**94%** (worst)

Key finding: All AI search engines collectively provided incorrect answers to more than 60% of queries. Even Perplexity, the best performer, was wrong more than 1 in 3 times.

Source: Columbia Journalism Review / Tow Center — AI Search Has a Citation Problem; Nieman Journalism Lab

Perplexity's Self-Reported Benchmarks

Perplexity publishes strong numbers from its own testing:

Metric	Score
SimpleQA factual accuracy	93.9%
Citation precision (when it cites, does it match?)	99.98%
Deep Research on Humanity's Last Exam	21.1% (beat OpenAI o3, Gemini Thinking, DeepSeek-R1)

Important caveat: SimpleQA tests simple factual recall ("What year did Python 3.0 release?"), not the nuanced citation-accuracy tasks the CJR study evaluated. Self-reported benchmarks and independent testing often tell very different stories.

Source: Perplexity — Introducing Deep Research

The "Second-Hand Hallucination" Problem

GPTZero's CEO Edward Tian uncovered a critical flaw in Perplexity's approach: because Perplexity searches the web for sources, and the web increasingly contains AI-generated content, Perplexity ends up **citing AI-generated errors as if they were reliable sources**.

Key findings:

- The average user encounters an **AI-generated source after just 3 prompts** to Perplexity
- In one documented case, Perplexity cited a single AI-generated LinkedIn article as its sole source

- This creates an **AI misinformation loop**: AI-generated errors on the web get cited by Perplexity, which generates AI-spun answers containing those same errors

This is particularly relevant for niche topics (like hearing aid specifications or Australian healthcare policy) where there may be fewer high-quality human-written sources and more AI-generated content filling the gap.

Source: GPTZero — Second-Hand Hallucinations: Investigating Perplexity's AI-Generated Sources; Futurism — Perplexity Is Already Citing Error-Filled AI-Generated Spam

An Ironic Twist: Paid Versions Perform Worse

The CJR study found a counterintuitive result: **paid versions of AI search tools (including Perplexity Pro at $20/month) provided more incorrect answers than free versions.** This is likely because Pro modes attempt more complex multi-step reasoning, which introduces more opportunities for error.

Perplexity Plans and Pricing (2026)

Plan	Price	Best For
Free	$0	Daily lookups, quick research — handles most needs
Pro	$20/month	Regular research, model selection, Deep Research reports
Max	$200/month	Heavy users, all frontier models, Model Council (runs 3 models simultaneously)

The **Model Council** feature (Pro and above) runs your query through three different AI models and compares their outputs — essentially a built-in cross-checking mechanism.

How a Hearing Clinic Owner Could Use Perplexity

Clinical research:

- "Compare the latest Phonak Lumity vs Oticon Intent for moderate sensorineural hearing loss" — gets cited answers from manufacturer specs and audiological reviews
- "Which common medications are ototoxic and what monitoring is recommended?" — with medical citations

Business operations:

- "What are the current NDIS hearing services guidelines for adults in Australia 2026?" — retrieves government policy documents
- "Current pay rates for audiologists under the Health Professionals and Support Services Award 2026"

Marketing and admin:

- "What are the most effective digital marketing strategies for audiology practices in regional Australia?"
- "How does GST apply to hearing aid sales in Australia?"

The Critical Rule: Always Click the Sources

Perplexity's numbered citations are its greatest strength — and they're only useful if you click them. For anything you'll act on (clinical decisions, regulatory compliance, financial matters), **open the source link and verify the claim directly**. The 37% failure rate means roughly 1 in 3 answers may contain inaccuracies. Treat Perplexity as a research accelerator — a fast way to find relevant sources — not as a final authority.

Sources: CJR / Tow Center Study; GPTZero Investigation; Perplexity Pricing

---

Microsoft Copilot: The One You Probably Already Have

If your clinic runs Microsoft 365 (Word, Excel, Outlook, Teams), you may already have Copilot — or been offered an upgrade that includes it. This makes Copilot the most likely AI tool for many hearing clinic owners to encounter, whether they actively chose it or not.

How Copilot Differs From Other AI Tools

Copilot is built on OpenAI's GPT models (currently GPT-4 Turbo) but adds two layers on top:

1. **Bing search integration** (for the free web version) — similar to Perplexity's approach
2. **Microsoft Graph grounding** (for the business version) — searches your own emails, documents, Teams messages, and SharePoint files before generating answers

This RAG-like architecture means Copilot in Microsoft 365 is designed to answer questions grounded in *your organisation's data*, not just the open web.

Copilot's Hallucination Performance

CJR / Tow Center Study (March 2025):

Copilot stood out in this study — but not in the way Microsoft would want. It **declined to answer 104 out of 200 queries** (52%). Of the remaining 96 it did attempt:

Result	Count	Percentage
Completely correct	16	17%
Partially correct	14	15%
Completely incorrect	66	**69%**

So when Copilot *did* answer, it was wrong roughly **70% of the time** on citation accuracy tasks. However, its high refusal rate means it was at least honest about its limitations more often than Grok or Gemini, which confidently gave wrong answers.

Source: CJR / Tow Center Study

Documented Copilot Hallucination Issues

In enterprise and consumer testing, Copilot has exhibited several specific hallucination patterns:

- **False accusations:** Copilot falsely identified a court reporter as an escaped psychiatric patient and convicted child abuser — he had only *reported* on these crimes
- **Word hallucinations:** Copilot in Word cannot access the web but doesn't tell users this, instead fabricating quotes, statistics, and citations when asked for them
- **Sales document fabrications:** In testing, Copilot invented product names that didn't exist and cited benefits that hadn't been requested
- **Business data misinterpretation:** Enterprise users reported erroneous document summaries, statistics that don't add up, and suggestions not supported by the organisation's actual data

Sources: The Conversation — Why Copilot Falsely Accused a Court Reporter; Computerworld — Is Copilot a Lying Liar?; Microsoft Community Hub — Hallucinations in Word

Relevant for Australian Users: The ACCC Lawsuit

In October 2025, the Australian Competition and Consumer Commission (ACCC) sued Microsoft for allegedly misleading approximately **2.7 million Australian customers** when integrating Copilot into Microsoft 365 plans. The ACCC alleges Microsoft told subscribers they had to accept Copilot (and pay higher prices) or cancel — while hiding the option to keep their existing plan without Copilot at the original price.

The Personal plan price increased **45% to A$159** and the Family plan rose **29% to A$179** following Copilot integration.

This is particularly relevant for clinic owners who may have had Copilot added to their Microsoft 365 subscription without fully understanding what changed — or being given a genuine choice.

Source: ACCC Media Release

Copilot Pricing and Plans

Plan	Price	Key Features
Free (Copilot)	$0	Built into Windows 11, Edge, Bing. GPT-4 Turbo, web search
Copilot Pro	USD $20/month	Priority model access, Microsoft 365 app integration, voice
M365 Copilot Business	USD $21/user/month	Full integration with Word, Excel, Outlook, Teams. Microsoft Graph grounding. Enterprise security

The M365 Copilot Business plan (launched December 2025) is the one most relevant for clinics — it works inside your existing Office apps and uses your own business data. A 15% introductory discount is available until March 31, 2026.

Source: Microsoft 365 Copilot Pricing

How a Hearing Clinic Could Use Copilot

Because Copilot lives inside tools you already use, it's particularly practical:

- **Outlook:** "Summarise the email thread with the GN Hearing rep" or "Draft a reply to this patient reschedule request"
- **Word:** "Create a referral letter template for GPs based on our existing letters" (but verify it doesn't fabricate details)
- **Excel:** "Analyse this month's appointment data and identify no-show patterns"
- **Teams:** Auto-summarise meeting notes and action items from staff meetings

Microsoft's "Usefully Wrong" Position

It's worth noting Microsoft's own framing. In its introductory Copilot blog, Microsoft stated: *"Sometimes Copilot will be right, other times usefully wrong — but it will always put you further ahead."* The company positions Copilot as a "first draft" tool that users should edit and iterate on. Whether "usefully wrong" is an acceptable standard for business tools is a question every clinic owner should consider.

Dragon Copilot: A Healthcare-Specific Option

Microsoft also offers **Dragon Copilot**, a separate healthcare-specific product that could be relevant for audiologists:

- Real-time dictation of clinical notes during patient consultations
- Captures clinician-patient conversations to generate draft medical notes
- Surfaces relevant patient history without digging through records

This is a specialised product (not included in standard M365 Copilot) but worth investigating for clinics looking to reduce admin burden.

Source: Microsoft — Dragon Copilot Use Cases

The Critical Warning

Copilot in Word is especially prone to hallucination because it **cannot access the web but doesn't tell you**. If you ask it to include statistics, quotes, or citations, it will confidently fabricate them. Always verify any factual claims Copilot adds to your documents, particularly any statistics, references, or regulatory information.

Grok: The Curious Outlier

Grok, built by Elon Musk's xAI, is the most polarising AI tool available. It produces some of the most fascinating — and contradictory — benchmark results of any model, and comes with significant controversy. You're unlikely to encounter it in a professional hearing clinic setting, but it's worth understanding why.

Grok's Contradictory Benchmark Story

Grok's hallucination data reads like it can't be about the same product:

Benchmark	Grok Result	What It Means
CJR Study (March 2025)	Grok-3: **94% failure rate** (worst of 8 tools)	When searching for news citations, almost every answer was wrong
Relum Study (Dec 2025)	Grok: **8%** (best of 10 models)	Workplace questions — but see credibility note below
Vectara Leaderboard	Grok-3: **5.8%**	For document summarisation, reasonably good
Vectara Leaderboard	Grok-4 (fast): **17.8–20.2%**	Newer versions actually scored *worse*

How can the same product be simultaneously the best and worst? The answer is the same lesson we saw with Claude: **the benchmark determines the story**. Grok-3 is genuinely strong at summarising provided documents (5.8% on Vectara) but catastrophically bad at searching the web for accurate citations (94% failure, generating 154 broken URLs out of 200 attempts in the CJR study).

A critical caveat on the Relum study: The widely-cited 8% hallucination figure comes from Relum, a *casino games aggregator platform* — not an AI research organisation, university, or established benchmarking body. The methodology has not been peer-reviewed, and the study was heavily amplified by pro-Musk media. It should not be weighted equally with the CJR study (Columbia University) or the Vectara leaderboard (standardised, open-source methodology).

Sources: CJR / Tow Center Study; Vectara Leaderboard (GitHub)

The Safety Factor

Unlike ChatGPT, Claude, and Gemini — which invest heavily in content safety guardrails — Grok has taken a deliberately "unfiltered" approach. This has led to several documented misinformation incidents, including providing incorrect election information to millions of users during the 2024 US presidential race that took over a week to correct.

Independent security researchers have also found that Grok's safety systems are significantly weaker than its competitors. One analysis of Grok 4 concluded it has **"no meaningful safety guardrails"** and could be trivially prompted to generate harmful or misleading content.

Grok also offers a "Fun Mode" that relaxes safety constraints for more opinionated, sarcastic responses — entertaining, but a liability for professional use.

Sources: Axios — Grok Election Misinformation; LessWrong — Grok 4 Safety Analysis; Australian Institute of International Affairs — What the Grok Controversy Reveals

Who Actually Uses Grok?

Grok has approximately **30 million monthly active users** — compared to ~400 million each for ChatGPT and Gemini, and ~20-35 million for Claude. It's primarily available through X (formerly Twitter) Premium subscriptions ($40/month) or the SuperGrok standalone plan ($30/month). Its user base skews heavily toward X power users, content creators, and software developers rather than business professionals. India accounts for 33% of users; the US 14%; Australia is not in the top markets. The Australian eSafety Commissioner maintains a guidance page on Grok, indicating regulatory awareness but not widespread professional adoption.

There is **no evidence of significant use in Australian healthcare or professional services**.

The Bottom Line on Grok

Grok is genuinely interesting from a technology perspective — its strong performance on some benchmarks (5.8% on Vectara) shows the underlying models have real capability. But the combination of **catastrophic web search failures (94% error rate), documented misinformation incidents, and weaker safety guardrails** makes it unsuitable for business or healthcare use.

If you're curious about AI capabilities and want to experiment, Grok can be fascinating. If you're running a clinic and need reliable information, stick with ChatGPT, Claude, Perplexity, or Copilot.

Privacy and Patient Data: What Australian Hearing Clinics Must Know

AI accuracy isn't the only risk — privacy is equally important. As a health service provider, your obligations under Australian law are stricter than for most small businesses.

You Are Covered by the Privacy Act — Regardless of Revenue

The standard small business exemption (under $3 million turnover) **does not apply to health service providers**. The Office of the Australian Information Commissioner (OAIC) explicitly lists audiologists among the health professionals covered by the Privacy Act 1988. This means all 13 Australian Privacy Principles (APPs) apply to your clinic, including:

- **APP 6** — Health information may only be used for the purpose it was collected, or a directly related secondary purpose
- **APP 8** — Cross-border disclosure requires reasonable steps to ensure overseas recipients handle data in accordance with Australian privacy standards

- **APP 11** — You must take reasonable steps to protect health information from misuse, interference, loss, and unauthorised access

If you're in Victoria, the Health Records Act 2001 adds further obligations. NSW has the Health Records and Information Privacy Act 2002. Other states rely on the Commonwealth Privacy Act.

Source: OAIC — Guide to Health Privacy (May 2025)

What Happens When You Paste Patient Data into AI Tools?

This is the critical question. Most free-tier AI tools **train on your conversations by default**:

Tool	Free Tier: Trains on Data?	Safe Tier (No Training)	Approximate Cost
ChatGPT	Yes (opt-out available)	Team or higher	$25–30/seat/month
Claude	User choice (opt-in/out)	Team or higher, or API	Variable
Gemini	Yes (conversations may be human-reviewed)	Workspace Business/Enterprise	Bundled with Workspace
Copilot	May use data for improvements	Microsoft 365 Copilot	Bundled with M365
Perplexity	Yes (opt-out available)	Enterprise or API	Custom pricing

Even with "safe" tiers, pasting identifiable patient data means the data is processed on overseas servers (typically US). Under APP 8, you must take reasonable steps to ensure the overseas recipient handles the information consistently with the APPs — which is difficult to guarantee with AI tools.

Important nuances:

- ChatGPT Free: If you give feedback (thumbs up/down), the **entire conversation** may be used for training, even with the training opt-out enabled
- Claude: Anthropic reduced API data retention to 7 days in September 2025; consumer plans retain data for 30 days (or up to 5 years if you opt in to training)
- Gemini: The consumer version is the riskiest — conversations may be reviewed by human evaluators. The Workspace-integrated version has enterprise protections

Sources: OpenAI Data Controls FAQ; Anthropic Privacy Center; Google Workspace AI Privacy Hub

AHPRA and Audiology Australia Guidance

AHPRA (August 2024) published five principles for AI use in healthcare:

1. **Accountability** — You remain responsible for any AI output used in your practice
2. **Understanding** — You should understand how the AI tools you use work, including their limitations
3. **Transparency** — Patients should be informed about AI use in their care

4. **Informed consent** — Patients must consent to AI use and their concerns must be addressed
5. **Ethical and legal implications** — Consider privacy obligations, data security, and professional ethics

Audiology Australia (September 2024) released a position statement on AI scribes, emphasising that using AI to handle sensitive patient data requires client consent, that audiologists must verify all AI-generated clinical notes, and that AI scribes are "in their infancy" and prone to errors including misinterpreting speech and clinical terminology.

Sources: AHPRA — AI in Healthcare Code of Conduct; Audiology Australia — AI Scribes Position Statement

The Regulatory Environment Is Tightening

Three recent developments signal increased accountability:

1. **Australian Clinical Labs — $5.8 million penalty (October 2025).** The first civil penalty under the Privacy Act, for failing to protect health information of 223,000 individuals. The health sector leads all industries for data breach notifications (18% of all breaches in the first half of 2025).
2. **Automated Decision-Making disclosure (from December 2026).** Under the Privacy and Other Legislation Amendment Act 2024, clinics must disclose in their privacy policies how AI tools are used in decisions that substantially affect individuals — including any AI-assisted clinical decisions.
3. **Proposed removal of the small business exemption (Tranche 2, timeline uncertain).** While hearing clinics are already covered, this would bring suppliers, IT providers, and contractors under the Privacy Act — potentially affecting your entire supply chain.

Sources: OAIC — ACL Penalty Media Release; Norton Rose Fulbright — Privacy Reform Analysis

Practical Privacy Guidelines

Do	Don't
De-identify all patient information before pasting into AI (use "Male, 70s, bilateral moderate SNHL" instead of full details)	Paste patient names, DOBs, Medicare numbers, or addresses into any AI tool
Use business/enterprise tiers for regular AI use	Rely on free-tier opt-outs for patient-related work
Add an AI disclosure to your patient intake forms	Use AI scribes or clinical note-taking tools without informed consent
Review and verify all AI-generated clinical content	Accept AI-generated clinical notes without checking
Train staff on which AI tools are approved and how to de-identify data	Assume staff know what's acceptable

Do	Don't
Keep a record of which AI tools your clinic uses	Ignore the December 2026 automated decision-making disclosure deadline
Opt out of training data on all consumer AI tools	Click thumbs up/down on ChatGPT conversations containing any sensitive information

Key Evaluation Frameworks and Benchmarks

For those who want to go deeper, these are the major benchmarks used to measure AI hallucination rates:

Benchmark	What It Measures	Who Runs It
Vectara HHEM Leaderboard	Factual consistency in document summarisation (98 models)	Vectara (via Hugging Face)
Galileo Hallucination Index	Overall accuracy across Q&A, RAG, and long-form tasks (22 models)	Galileo AI
HalluLens	Extrinsic hallucination across 3 task types with dynamic test sets	ACL 2025 academic paper
TruthfulQA	Whether models generate truthful answers vs common misconceptions	Academic benchmark
FActScore	Fine-grained factual precision in long-form text generation	Academic benchmark

Sources: Vectara Leaderboard (Hugging Face); Galileo Hallucination Index; HalluLens (ACL 2025)

The Bottom Line

AI hallucinations are a real and measurable phenomenon, but they are **not a reason to avoid AI** — they're a reason to use it smartly. The data tells a clear story:

1. **AI is getting better fast.** Grounded task hallucination rates have dropped from ~15% to under 1% in three years.
2. **Context is everything.** Give AI source material, and accuracy soars. Ask it to recall facts from memory, and risks increase.
3. **Benchmarks don't tell the whole story.** A model can rank #1 on one benchmark and bottom-tier on another — because they measure different things. Always ask *what* is being

measured.

4. **Simple habits make a big difference.** Using structured prompts (like RACE), asking for sources, and spending 10 seconds reviewing output catches nearly all errors.

5. **AI is a tool, not a truth-teller.** Just like you'd proofread an email or double-check a financial calculation, you review AI output — especially for anything published, sent to clients, or involving professional advice.

The businesses that will benefit most from AI aren't the ones that trust it blindly or avoid it entirely. They're the ones that **use it wisely, verify what matters, and save hours every week** on the tasks where AI genuinely excels.

Sources and Further Reading

Benchmarks and Leaderboards

Vectara Hallucination Leaderboard (Hugging Face) — 98 models ranked, updated regularly

Vectara GitHub — Hallucination Leaderboard

Galileo Hallucination Index — 22 models evaluated across multiple task types

Artificial Analysis — AA-Omniscience Benchmark — Factual knowledge + abstention calibration

HalluLens: LLM Hallucination Benchmark (ACL 2025) — Academic benchmark with dynamic test sets

Key Studies and Reports

Hallucinating Law: Legal Mistakes with LLMs (Stanford, 2024)

Hallucination-Free? Assessing AI Legal Research Tools (Stanford, 2025)

GPTZero: 100 Hallucinations in NeurIPS 2025 Papers

NeurIPS Hallucinated Citations (Fortune, Jan 2026)

NeurIPS Hallucinated Citations (TechCrunch, Jan 2026)

Columbia Journalism Review — AI Search Engine Accuracy (March 2025)

Foundational Reading

Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Meta AI, 2020 (RAG founding paper)

AI Hallucination Statistics 2026 (All About AI)

Hallucination Rates in 2025 (Markus Brinsa, Medium)

Key Research Papers

OpenAI — Why Language Models Hallucinate (Sept 2025) — OpenAI's own analysis of why hallucinations are structural

OpenAI's Reasoning Models Hallucinate More (TechCrunch, April 2025)

FaithBench: A Diverse Hallucination Benchmark for Summarization (arXiv, NAACL 2025) — Found 21.3% "benign hallucination" rate in Claude summaries

Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards (arXiv, 2025) — Found 16 ranking inversions in HHEM vs human judgements

HalluHard: A Hard Multi-Turn Hallucination Benchmark (2026)

Hallucination Mitigation in RAG, Reasoning, and Agentic Systems (arXiv, 2025)

Nature — Taxonomy of AI Hallucinations (2024)

Perplexity-Specific Sources

AI Search Has a Citation Problem — Columbia Journalism Review / Tow Center (March 2025)

AI Search Engines Fail to Produce Accurate Citations in Over 60% of Tests — Nieman Journalism Lab

Second-Hand Hallucinations: Investigating Perplexity's AI-Generated Sources — GPTZero

Perplexity Is Already Citing Error-Filled AI-Generated Spam — Futurism

Britannica and Merriam-Webster Sue Perplexity Over Hallucinated Attributions (Sept 2025)

Introducing Perplexity Deep Research — Perplexity Blog

Copilot-Specific Sources

ACCC — Microsoft in Court for Allegedly Misleading Millions of Australians (Oct 2025)

The Conversation — Why Copilot Falsely Accused a Court Reporter of Crimes

Computerworld — Is Copilot for Microsoft 365 a Lying Liar?

Microsoft Community Hub — Hallucinations With Copilot in Word

Microsoft 365 Copilot Architecture — How It Works

Microsoft 365 Copilot Pricing

Grok-Specific Sources

Axios — Grok Spreads Election Misinformation (Aug 2024)

LessWrong — xAI's Grok 4 Has No Meaningful Safety Guardrails

Australian Institute of International Affairs — What the Grok Controversy Reveals

Claude-Specific Sources

Artificial Analysis — Claude Opus 4.6 Overview

Anthropic CEO: AI Models Probably Hallucinate Less Than Humans (TechCrunch, May 2025)

Why Claude Refuses 45% of Questions (B2B News Network, Dec 2025)

Reduce Hallucinations — Claude Documentation

Privacy and Regulatory Sources

OAIC — Guide to Health Privacy (May 2025)

OAIC — What is a Health Service Provider?

OAIC — Australian Clinical Labs $5.8M Penalty (Oct 2025)

AHPRA — AI in Healthcare Code of Conduct (Aug 2024)

Audiology Australia — AI Scribes Position Statement (Sept 2024)

ACSQHC — Pragmatic AI Guidance for Clinicians (Aug 2025)

Safer Care Victoria — ChatGPT and Generative AI Advisory (July 2023)

Norton Rose Fulbright — Privacy and Other Legislation Amendment Act 2024 Analysis

OpenAI Data Controls FAQ

Anthropic Privacy Center — Data Retention

Google Workspace AI Privacy Hub

Perplexity Privacy & Security

Microsoft 365 Copilot Enterprise Data Protection

News Coverage of Incidents

Deloitte AI Report Scandal (Fortune, Oct 2025)

Deloitte AI Debacle — Wake-Up Call (CFO Dive)

Deloitte Refund Details (CFO Dive)

AI Models Least & Most Likely to Invent Information (TechRepublic)

Stanford Study: AI Legal Research Tools Prone to Hallucinations (VentureBeat)

About Hearpreneur Solutions

Hearpreneur Solutions helps hearing clinic owners build thriving, values-driven practices. We provide business consulting, compliance support, and practical tools tailored specifically for the Australian hearing industry.

Want to discuss how AI can work for your clinic?

Book a free discovery call at **www.hearpreneur.com.au**

Article prepared February 2026. AI hallucination rates and benchmarks are rapidly evolving — always check the latest leaderboard data for the most current figures.

This article was researched using AI tools with human verification of all claims, benchmarks, and sources — practising what it preaches.

HEARPRENEUR
SOLUTIONS
THRIVE WITH VALUES