Hallucination Defence Kit

Stop AI from making things up in your work.
The workflow, the prompts, the toggles.

> *Hallucinations are mathematically inevitable with current training methods. The fix is your workflow.*
>
> - OpenAI, Why Language Models Hallucinate, Sept 2025

92% of people never check what their AI tells them. The model is 34% more confident when it's wrong than when it's right. This kit is the defence workflow I use before publishing anything AI has touched.

Five sections. One checklist. One prompt template. Use them every time.

By Kyle Balmer
AI with Kyle | aiwithkyle.com

What's Actually Happening

Large language models do not look anything up. There is no database. They predict the next most probable token, then the next, then the next, based on patterns from training. When that prediction lands on something true, we call it intelligent. When it doesn't, we call it a hallucination. Same mechanism either way.

Two flavours:

Type 1

Invents something that does not exist.

Fake book. Fake court case. Fake statistic. Easier to spot.

Type 2

Drifts from a document you actually gave it.

Misquoted line. Changed number. Plausible. Much harder to spot.

The Numbers

92%	of people never verify their AI answers
45%	of AI responses have significant problems (BBC/EBU, 3,000 tested)
34%	more confident when wrong than when right (MIT, Jan 2025)
486+	court cases globally with AI-fabricated citations

A liar knows the truth and hides it. AI has no concept of truth at all. It can't lie. It's a stochastic parrot. Once you internalise this, you stop trusting AI in the wrong places and start using it in the right ones.

When Hallucinations Spike

Four situations where hallucination risk goes through the roof. Slow down in any of them.

Specific facts

Citations, statistics, names, dates, quotes. The more precise it sounds, the more worth checking. Famous fake quotes (Einstein on insanity, Mandela on education, Marilyn Monroe on imperfection) fool models constantly.

Niche topics

The rarer the subject, the less training data. The model fills the gap. Same goes for less-represented languages - English is 26%% of internet content, Chinese 19%%, Spanish 8%%, Arabic 5%%, then it falls off a cliff.

Recent events

Every model has a training cutoff. Anything after it, it doesn't know. Good models realise and reach for search. Free-tier models often confidently make something up. Always check the dates.

Wrong assumptions

AI tends to agree with your premise. Ask 'is this a good idea?' and ChatGPT will say yes. Ask a leading question, get a led answer. The voice mode in particular is a relentless yes-man.

The Four-Toggle Setup

Before you ask a high-stakes question, run through this list. Most people miss two or three of these. The cumulative effect is dramatic.

1. Web search ON

ChatGPT, Claude, Gemini all have a toggle. On free plans it's often off by default. Flip it on. GPT-5 with web search makes ~45% fewer factual errors than GPT-4o on certain question types (OpenAI, GPT-5 System Card, Aug 2025).

2. Stronger model selected

Claude Opus 4.6/4.7 for high-stakes work. GPT-5 Pro with thinking on. Gemini 2.5 Pro. They're slower and cost more, but when accuracy matters they're worth it. The free models are fine for casual queries, not for client work.

3. Deep research / extended thinking ON

Every chatbot has its own name for this. ChatGPT calls it Deep Research. Claude calls it Extended Thinking. Gemini has Deep Think. Toggle it on for anything where the answer needs to be defensible, not just fast.

4. Source documents loaded

If you have the source material, give it to the model. Don't ask it to source from memory. NotebookLM is the easiest way - upload PDFs, transcripts, papers, then ask questions. Citation hallucination drops dramatically when grounded.

Quick check before you click send: 4 toggles on?

Web search. Stronger model. Deep thinking. Source loaded.

If you can't tick all four, slow down before you trust the answer.

Three Prompt Patterns

Small changes to how you ask. Massive change in what you get back. These are the three patterns I use every day.

Pattern 1: Source the question

Don't ask for a stat. Give it the source.

Risky: `"What's the stat on UK water usage?"`

Better: `"Here's the report. Pull the key stat and cite the exact line."`

Pattern 2: Force comparison

Don't ask is this a good idea. Give two options.

Risky: `"Is this a good business idea?"`

Better: `"Here are my two ideas. Which is stronger and why?"`

Pattern 3: Permit abstention

Add this single line to any prompt.

Add: `"If you're not certain, say so."`

Why: `models are trained to answer, abstaining gets penalised. Override that.`

Pre-Publish Checklist

Before any AI-touched output goes public - newsletter, client email, blog post, social, slide deck, anything with your name on it - run this list. If you can't tick all of these, don't publish yet.

- [ ] Web search was ON when the answer was generated
- [ ] Stronger model used (Opus / GPT-5 Pro / Gemini 2.5 Pro), not the free tier
- [ ] Every specific number, name, date, quote has been independently verified
- [ ] Every URL works and points to the page actually claimed
- [ ] Every cited paper, book, court case, study has been confirmed to exist
- [ ] Document drift checked: does the AI's summary match what the source actually says?
- [ ] No leading questions in the prompt chain (re-read the prompt, look for assumptions)
- [ ] If the model sounded very confident, that section was checked twice
- [ ] Anything outside the model's training cutoff was sourced fresh
- [ ] If publishing claims, you'd be willing to defend each one in court

The Deloitte rule: if a Big Four firm with billable consultants can ship two AI-hallucinated reports to two governments in one year, your team without that scrutiny is exposed. Run the list. Every time.

The Reframe

For 25 years the default way to get information online was retrieval.
Find the real thing. Show it to you.

That mental model does not apply here.

AI is a writing tool. A drafting tool. A thinking tool.

It is excellent at working with information you give it,
and unreliable at sourcing information from its own memory.

The people getting the most out of AI right now are not
treating it as a smarter search engine. They are treating
it as a brilliant drafter with an imperfect memory,
and they read the output before it goes anywhere.

Use AI to work with information. Don't trust it to source information from memory.

More from AI with Kyle

Daily newsletter: aiwithkyle.com
Live show: YouTube /@aiwithkyle (weekday mornings)
TikTok: @iamkylebalmer
Workshops: aiwithkyle.com/workshops

Sources: OpenAI, Why Language Models Hallucinate (Sept 2025); GPT-5 System Card (Aug 2025); BBC/EBU news AI study (Oct 2025); MIT confidence study (Jan 2025); Stanford RegLab, Large Legal Fictions (2024); Mata v. Avianca, 678 F. Supp. 3d 443 (S.D.N.Y. 2023).