

Hallucinations, Limitations and Critical Evaluation

Strategies for critically evaluating LLM responses

Alba Refoyo Martínez, PhD

Word of the year, 2023

hallucinate

“When an artificial intelligence hallucinates, it produces false information.”

Cambridge Dictionary · Dictionary.com chose the same word that year; Collins chose **AI**.

“**AI is incredibly smart and shockingly stupid.**”

Yejin Choi, TED2023

1

What a hallucination is

And why the mechanism produces them

2

Limitations

Non-determinism, ambiguity, context

3

Mitigation

Detection, uncertainty signals, RAG

4

Critical evaluation

What to do at your desk tomorrow

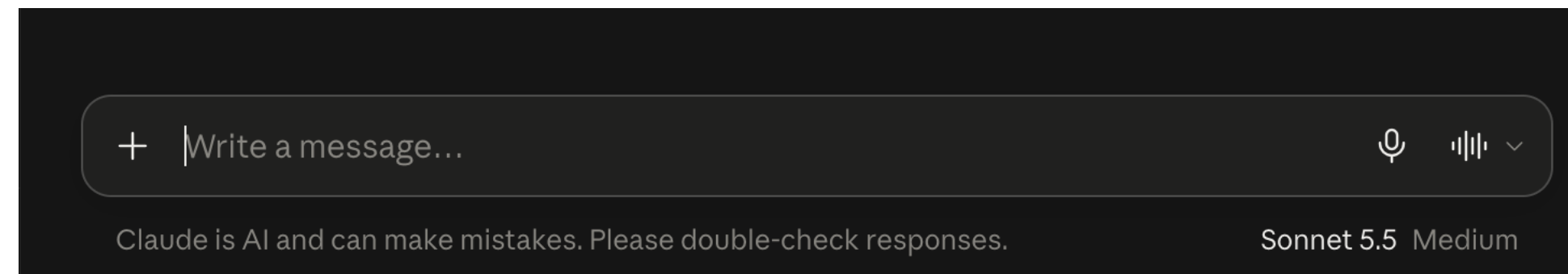
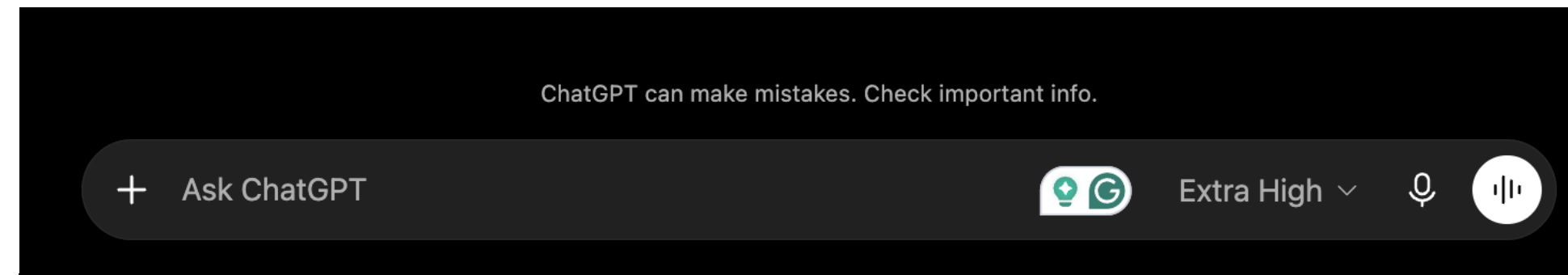
1

What a hallucination is

And why the mechanism produces them

Hallucination is **plausible**

Output that is fluent, grammatically correct, confident and wrong - contains inaccuracies or fabricated details.



Hallucination isn't some bug sitting in the code. It's a direct consequence of how the model was trained.

Spot the hallucination

Generated in response to “summarise what is known about metformin”. **Three claims here should stop you.** Which?

Metformin remains first-line pharmacotherapy for type 2 diabetes, acting primarily through inhibition of hepatic gluconeogenesis via AMPK activation. A 2019 randomised trial in *The Lancet* (Karlsson et al., 393:1502–1511) reported a 34.7% reduction in major cardiovascular events over 4.2 years of follow-up. The drug is contraindicated below an eGFR of 30 mL/min/1.73 m², and the *TCF7L2* variant on chromosome 11 is the strongest common genetic predictor of treatment response.

Nothing in the prose gives it away. True and fabricated claims are mixed in sentence by sentence.

When LLM's don't know, they make stuff up

Metformin remains first-line pharmacotherapy for type 2 diabetes, acting primarily through inhibition of hepatic gluconeogenesis via AMPK activation. A 2019 randomised trial in *The Lancet* (Karlsson et al., 393:1502–1511) reported a 34.7% reduction in major cardiovascular events over 4.2 years of follow-up. The drug is contraindicated below an eGFR of 30 mL/min/1.73 m², and the *TCF7L2* variant on chromosome 11 is the strongest common genetic predictor of treatment response.

A hallucination doesn't look like noise. It's internally consistent, correctly formatted, and written in the tone you expected, so it's **easy to miss and costly when you do**.

Two kinds of wrong

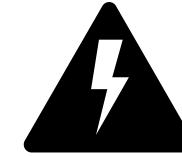
FAITHFULNESS · INTRINSIC

It contradicts the source you gave it.

Your paper says one thing; the summary says another.

Check: read it against your document

FACTUALITY · EXTRINSIC



It contradicts the world, with no source at all.

The invented citation. The wrong chromosome. Open-domain questions

Check: go out to PubMed, HGNC, the DOI, etc.

Which one you have decides which check works. Re-reading never catches the second kind (e.g. metformin).

Common examples

Fabricated research

Papers, authors, DOIs and citations that do not exist

Fake statistics

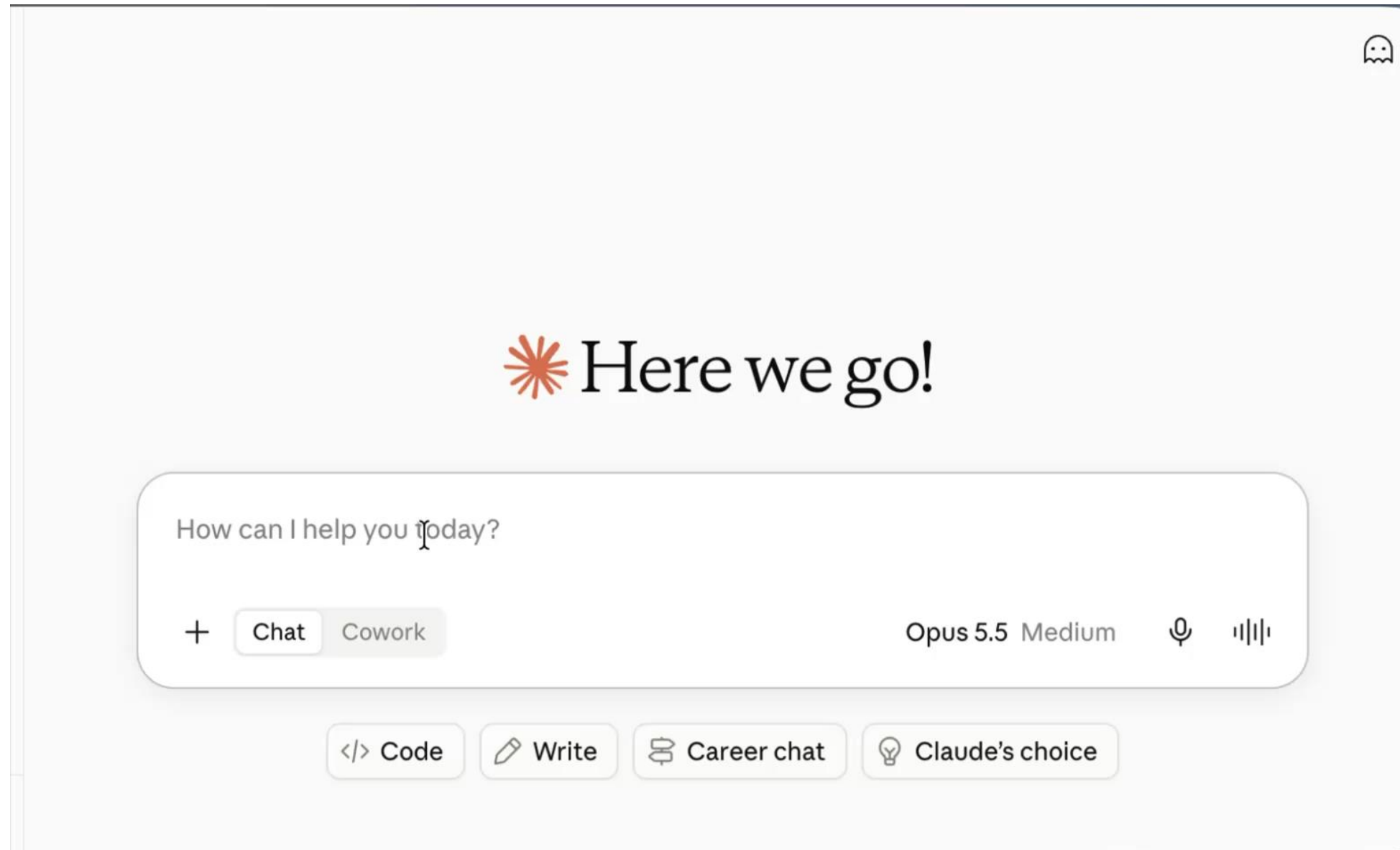
Precise-looking numbers, rates and percentages with no source

Wrong facts about people, events, etc.

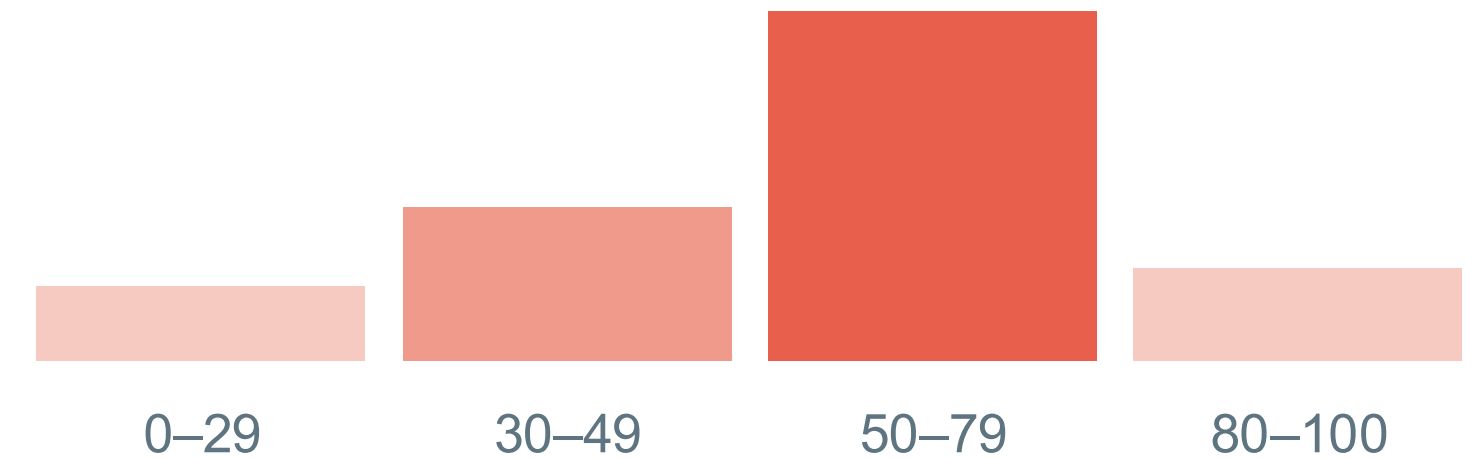
Invented quotes, dates, affiliations and outcomes

Metformin remains first-line pharmacotherapy for type 2 diabetes, acting primarily through inhibition of hepatic gluconeogenesis via AMPK activation. A 2019 randomised trial in *The Lancet* (Karlsson et al., 393:1502–1511) reported a 34.7% reduction in major cardiovascular events over 4.2 years of follow-up. The drug is contraindicated below an eGFR of 30 mL/min/1.73 m², and the *TCF7L2* variant on chromosome 11 is the strongest common genetic predictor of treatment response.

Ask any LLM for a random number



How many of you got a 42, 47 or 73?

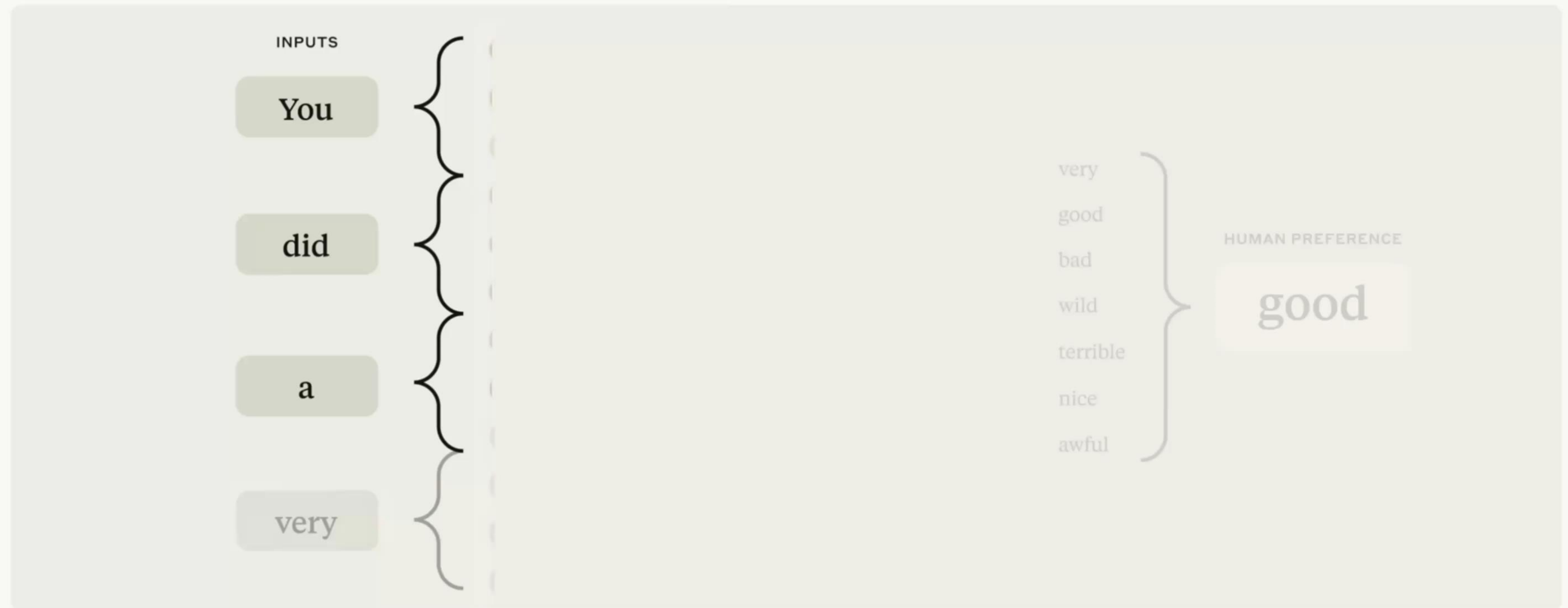


Raising the temperature widens the spread without making it uniform.

“Random” here just means whatever showed up most often in training. The model isn't sampling at random, it's recalling a pattern. **This is the same mechanism that generated the fabricated citation, just with a harmless output.**

Why they happen

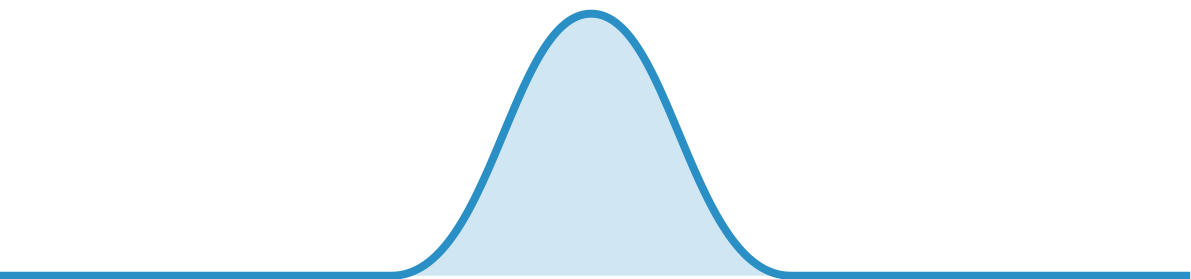
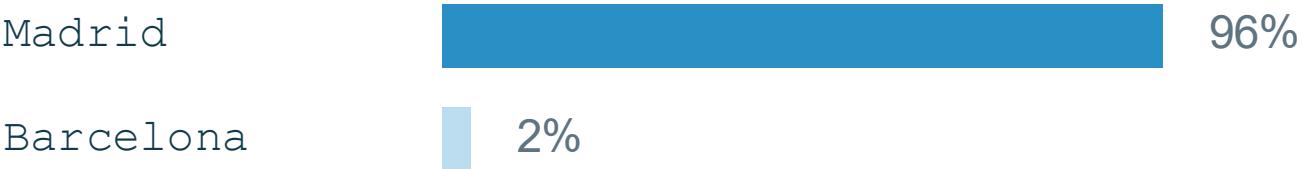
The model is calculating the **next most likely token**, learnt from training data. **Truth is not** a term in the **objective**.



Why they happen and what means in practice

SHARP DISTRIBUTION

The capital of Spain is ...

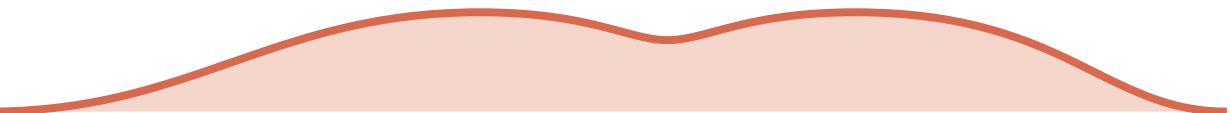
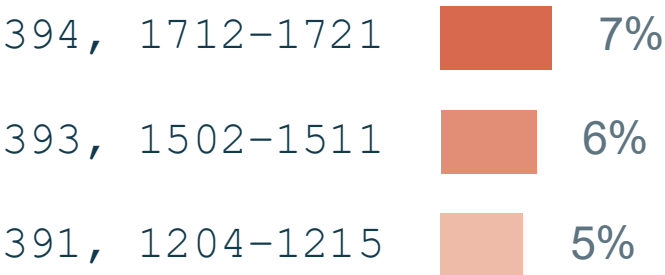


Frequent, well-attested facts are produced reliably because the distribution is sharp.

Training corpus contains the fact in this phrasing probably thousands of times.

FLAT DISTRIBUTION

The 2019 cohort study by Lindqvist et al., Lancet ...



It memorised the shape of a citation, not this one.

The model is just as fluent and confident producing a rare, shaky fact as a common, solid one.

When they happen

Asking for specifics

The topic is obscure, niche or very recent

It is real but not widely known

You need exact details

In this scenarios, LLM it takes a guess but that guess might be wrong. Nothing in cross-entropy loss rewards correctness as such, it rewards matching the distribution of the training data.

Where have you caught an LLM being confidently wrong?

And how did you notice?

Why they persist

Models are optimised to be good test-takers.

HOW THE BENCHMARK MARKS IT

	Correct answer	+1
	Confident guess, wrong	0
	“I don't know”	0

Under this scoring system, a guess always beats saying “I don't know.” It costs nothing extra compared with abstaining, but occasionally wins.

Confidence is what gets trained in.

Can we do something about it?



Computer Science > Computation and Language

[Submitted on 4 Sep 2025]

Why Language Models Hallucinate

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang

- ✗ **Generation is *harder (statistically)* than classification:** a model can judge a statement false yet still generate it. Pretraining on next-token prediction guarantees some baseline error rate no matter how good the data or architecture gets.
- ✓ It's not a model problem but a **grading problem**. Change how benchmarks score abstention, reward calibrated uncertainty, and the incentive to bluff goes away.

Accuracy: % correct

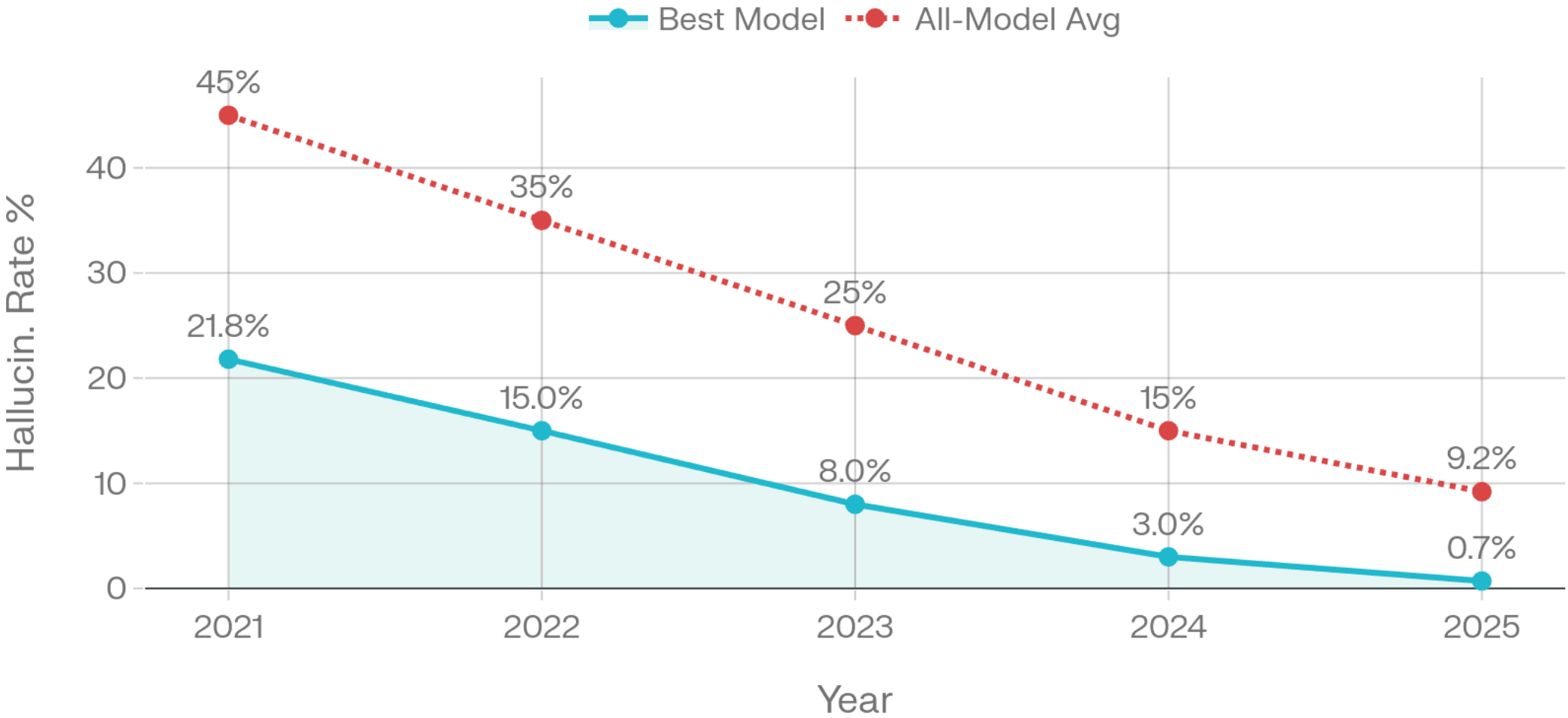
Hallucination rate: % guessed wrong

The hallucination rate specifically punishes bluffing rather than ignorance.

Rates really are falling

Hallucination Rates: 96% Drop in 4 Years

Source: AllAboutAI / Vectara | Summarization benchmark, best-in-class vs average



suprmind.ai/hub/ai-hallucination-rates-and-benchmarks

On **grounded summarisation**, best-in-class rates fell from 22% to under 1%.
Ground truth is in the prompt.

On **open-domain recall** the same models sit at **16 to 33%**.

Summarising a document you hand it is close to solved by now. Recalling a fact it was never taught is a different story.

THERE IS NO SINGLE HALLUCINATION RATE; THERE IS A RATE PER TASK.

How often do these models say I don't know?

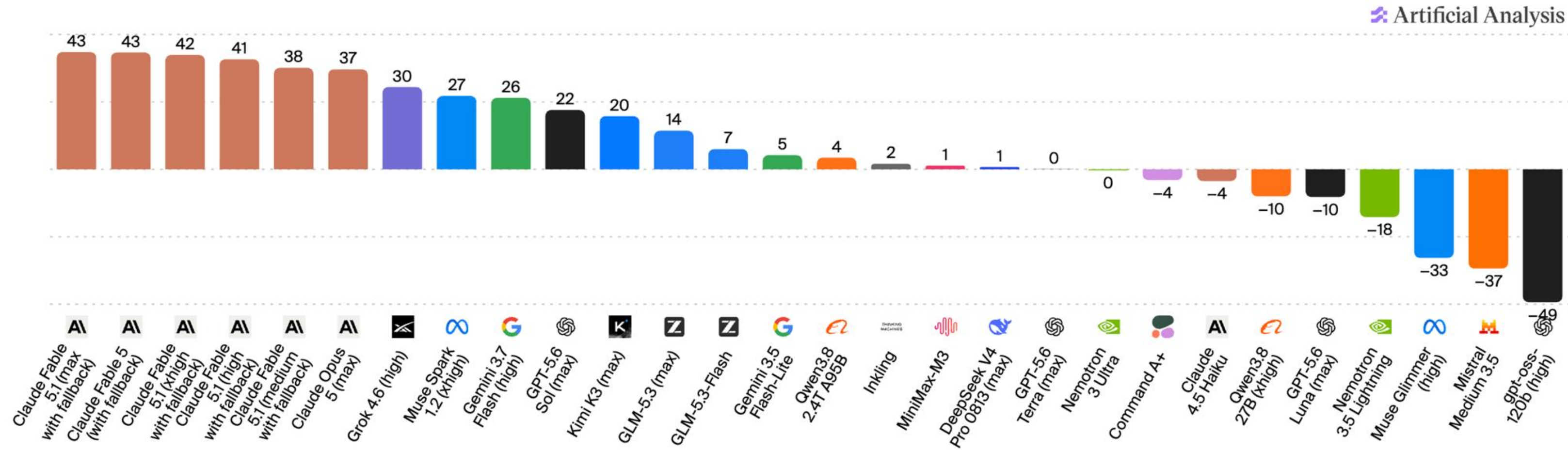
How often?

Omniscience Index:

- +1 correct,
- 1 wrong,
- 0 abstain

A model that says “I don’t know” scores better than one that guessed wrong.

AA-Omniscience: 6,000 questions, 42 topics, 6 domains, including health, software engineering and law.

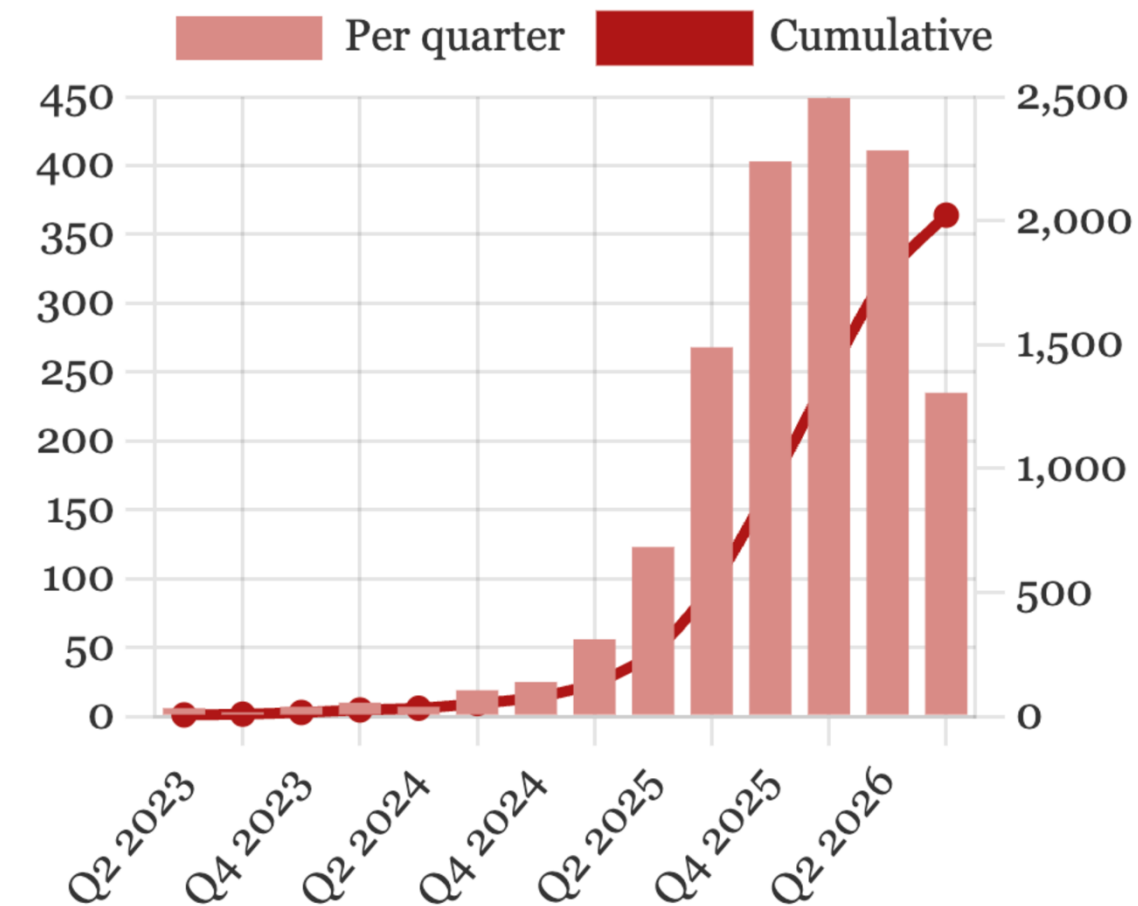


At launch, all but three models evaluated were more likely to hallucinate when they didn't know the answer:

What it costs in practice

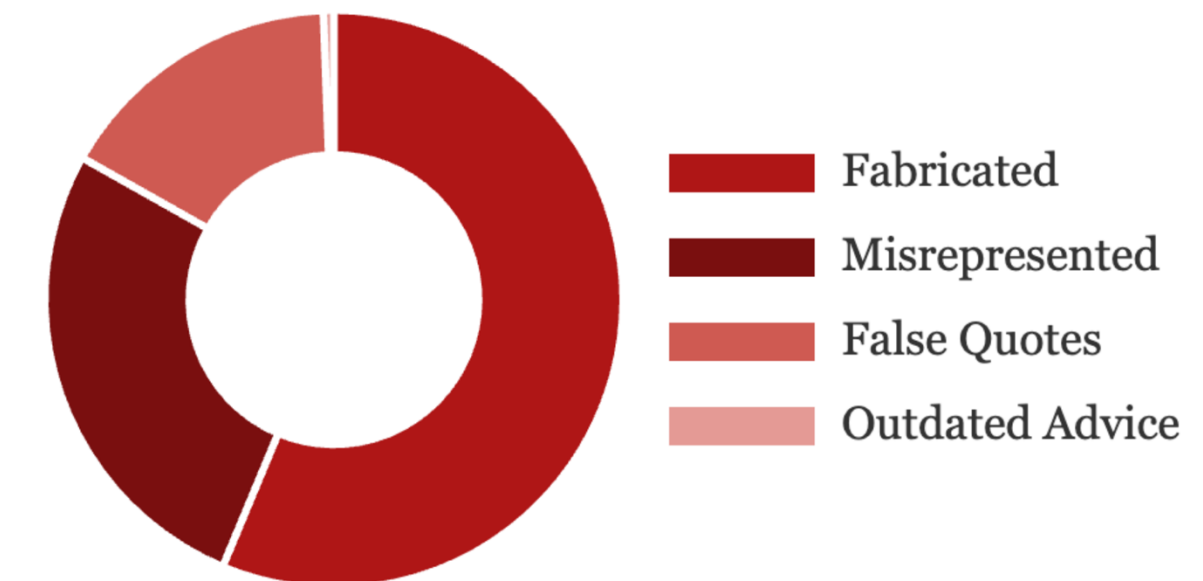
Database tracking court decisions in which a party relied on **hallucinated legal material**.

CASES OVER TIME (BY DECISION DATE)



n = 2022 decisions

NATURE OF HALLUCINATION



n = 6005 items across 1918 cases · 104 cases without items

These are trained professionals, working under oath, with every incentive to verify. They still missed it.

Bias vs. hallucination

Both read identically on the page. Neither is fixed by asking the model to be careful.

Trained on data scraped from the internet, which makes it well-trained but not necessarily well-balanced.

Consequence: the output **reflects the training**, so it reproduces whatever the corpus over-represents: English-language sources, Western clinical populations, US institutions.

BIAS

Bias is the model staying *faithful* to a skewed corpus.

HALLUCINATION

Hallucination is it staying faithful to nothing at all.

2

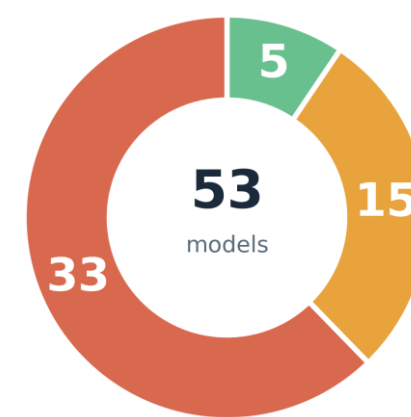
Limitations

What harnessing can and cannot reach

The car wash test

"I want to wash my car. The car wash is 50 meters away.
Should I walk or drive?"

One correct answer, one step of reasoning:
the car has to be at the car wash. 53 models,
530 calls. A human control group of 10,000
scored **71.5%**, beating 48 of the 53 models.



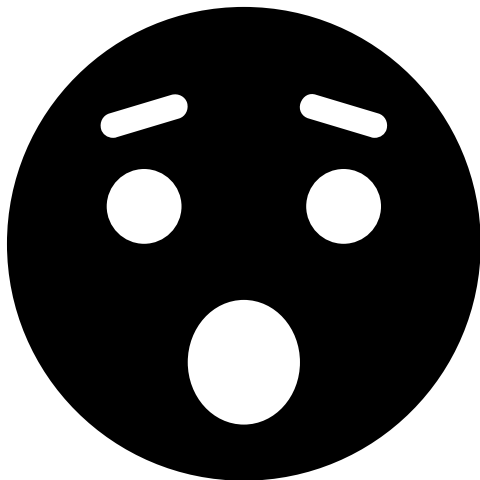
→ always right, 10 out of 10

→ sometimes right: the dangerous tier

→ never right

Limitation: a learned heuristic ("50 m is close → walk") overrides the stated goal. **Not a knowledge gap.**

A week ago,



I need to wash my car, a 100m away from my doorstep, do I walk or do I drive?

Weighing whether walking or driving suits a short errand.

Walk! Definitely walk.

Here's why:

1. **It's very short** — 100 meters is only about a 1-2 minute walk, so the walk time is negligible
2. **Driving is inefficient** — You'd spend more time getting your car ready and driving than you would just walking
3. **Easier with supplies** — You might need to carry a bucket, sponge, towels, or hose. Walking lets you make multiple trips if needed without the hassle of parking
4. **No practical advantage** — There's no meaningful benefit to driving such a short distance
5. **Better for the environment** — No emissions for such a minimal journey
6. **Good for you** — You get a quick bit of exercise

Save the car for longer distances. This is a perfect walking trip.





Write a message...


Haiku 4.5
Fastest for quick answers ✓

Effort Extended >

More models >

Haiku 4.5 Extended  

Claude is AI and can make mistakes. Please double-check responses.

Wrong answer, different causes

They read identically but rerunning the prompt tells them apart.

FAILURE	WHAT HAPPENED	RERUN IT FIVE TIMES	FIX (by you)
Ambiguity	The question had more than one reading, and the model committed to one without telling you.	Same answer every time, to a different question (not the one you asked)	State the goal, not just the task
Heuristic override	A strong pattern beat the reasoning. The capability exists but loses to a more frequent association.	Answers vary between runs	Decompose; ask for the constraint first
Fabrication	The fact was never there to recall. It's a flat distribution, sampled confidently.	A different fabrication every run	Ground it, or do not ask

Non-determinism

Same prompt, same settings, different answer. This is expected behaviour, not a fault.

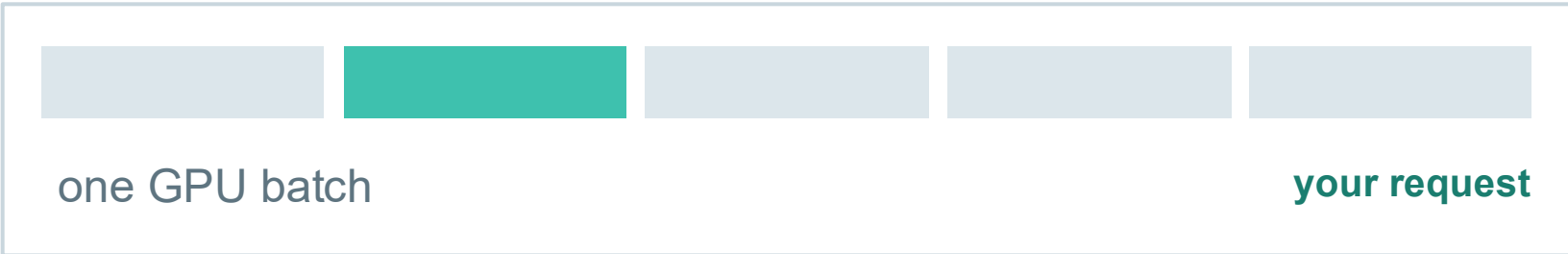
SAMPLING & TEMPERATURE

The model emits a distribution over next tokens; the sampler draws from it. At any temperature above zero the same prompt can take a different draw.

This is the one you control, `temperature=0`.

DYNAMIC BATCHING

The server batches your call with whoever else is in the queue. **You do not control this.**



E.g. The best city for a conference...

Batch shape changes the sum order, and that can flip a near-tie even at temperature 0.

Practical tuning

TEMPERATURE (0-1)

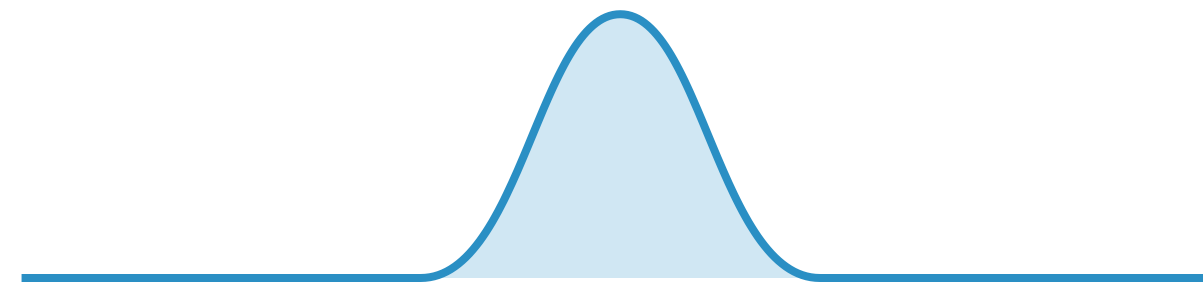
Controls how “adventurous” the pick is.

It reshapes the whole probability distribution (flatter == more random)

TOP-P (0-1)

Controls how many candidates are allowed

It cuts off the unlikely tail (keeping a small set of top tokens whose probability add up to at least p ; the rest are ignored)



What are you trying to get out of the model?

Practical tuning

Turn the temperature up and you get variety, not accuracy.

ACCURATE	BALANCED	NOVEL
TEMPERATURE 0 — 0.2	TEMPERATURE 0.5 — 0.7	TEMPERATURE 0.9 +
TOP-P 0.1 — 0.3	TOP-P 0.8 — 0.95	TOP-P 0.95 +
WHAT YOU GET The same answer twice	WHAT YOU GET Natural prose, wording moves but the content should not	WHAT YOU GET Diversity, creativity
USE IT FOR Extraction, code, methods text, structured output	USE IT FOR Summaries, drafting, rewriting your own prose	USE IT FOR Brainstorming, alternatives, framing options

Attention and long context

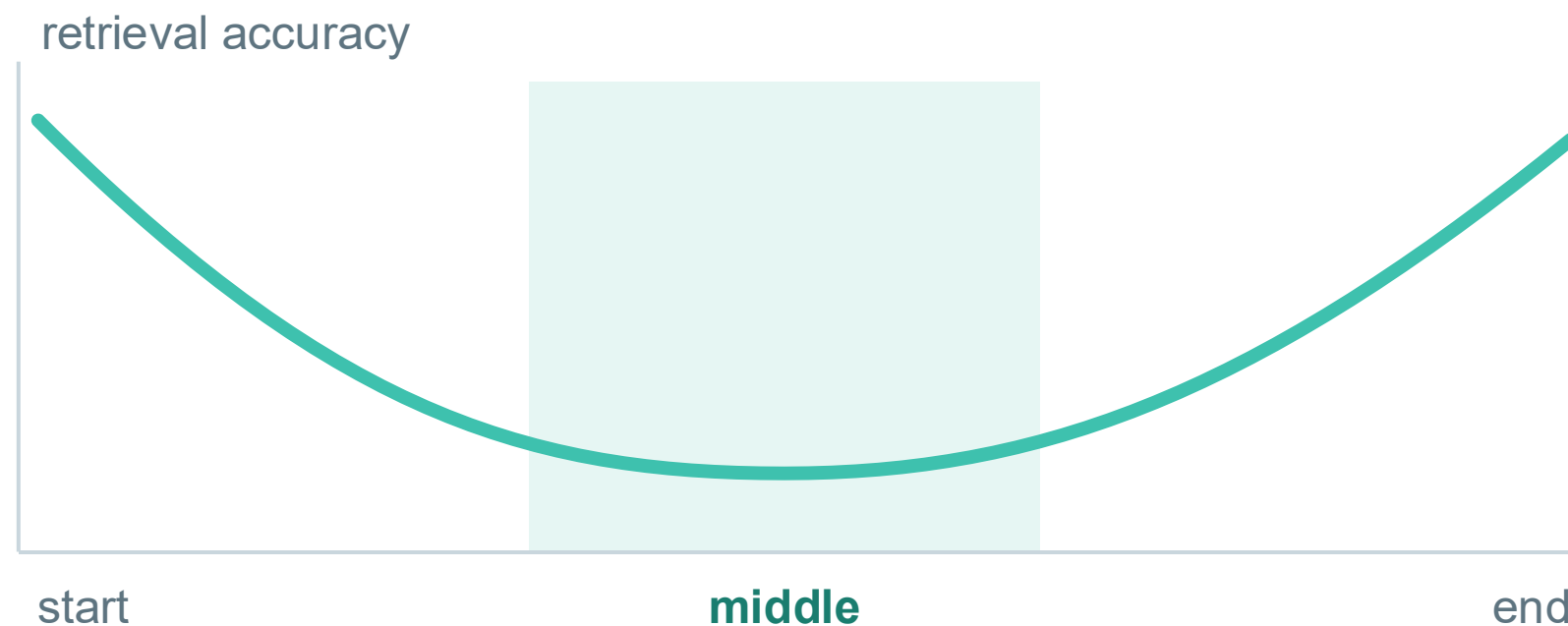
A large context window tells you about **capacity**, not attention. Retrieval accuracy is highest at the start and end of the input and degrades in the middle, in what's called the “lost in the middle” effect.

Pasting a 60-page protocol into the window does not mean the model read page 31.

The longer the conversation, the worse it gets: the model summarizes or drops earlier turns, and it never tells you which

WHAT TO DO INSTEAD

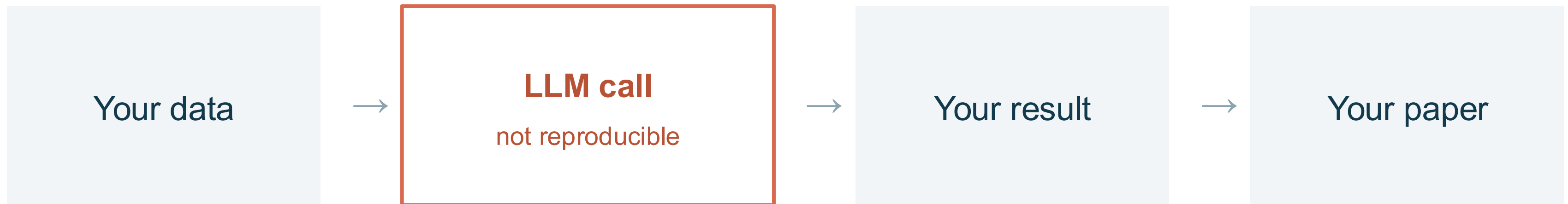
Chunk it. Ask one targeted question per chunk. Put the instruction **last**, not buried above the document.



Schematic, after Liu et al. (2024), *TACL* 12

An LLM call is not a reproducible step

IN A RESEARCH PIPELINE



Re-running the pipeline does not re-create that step.



```
2026-09-07 14:02 · claude-sonnet-4.6 · temp=0 · output saved → run_014.txt
```



Log the output, with model, version and date. That's the part worth keeping, not the prompt.

Can these limitations be mitigated?

What do you already do?

3

Mitigation

Detection · uncertainty · grounding

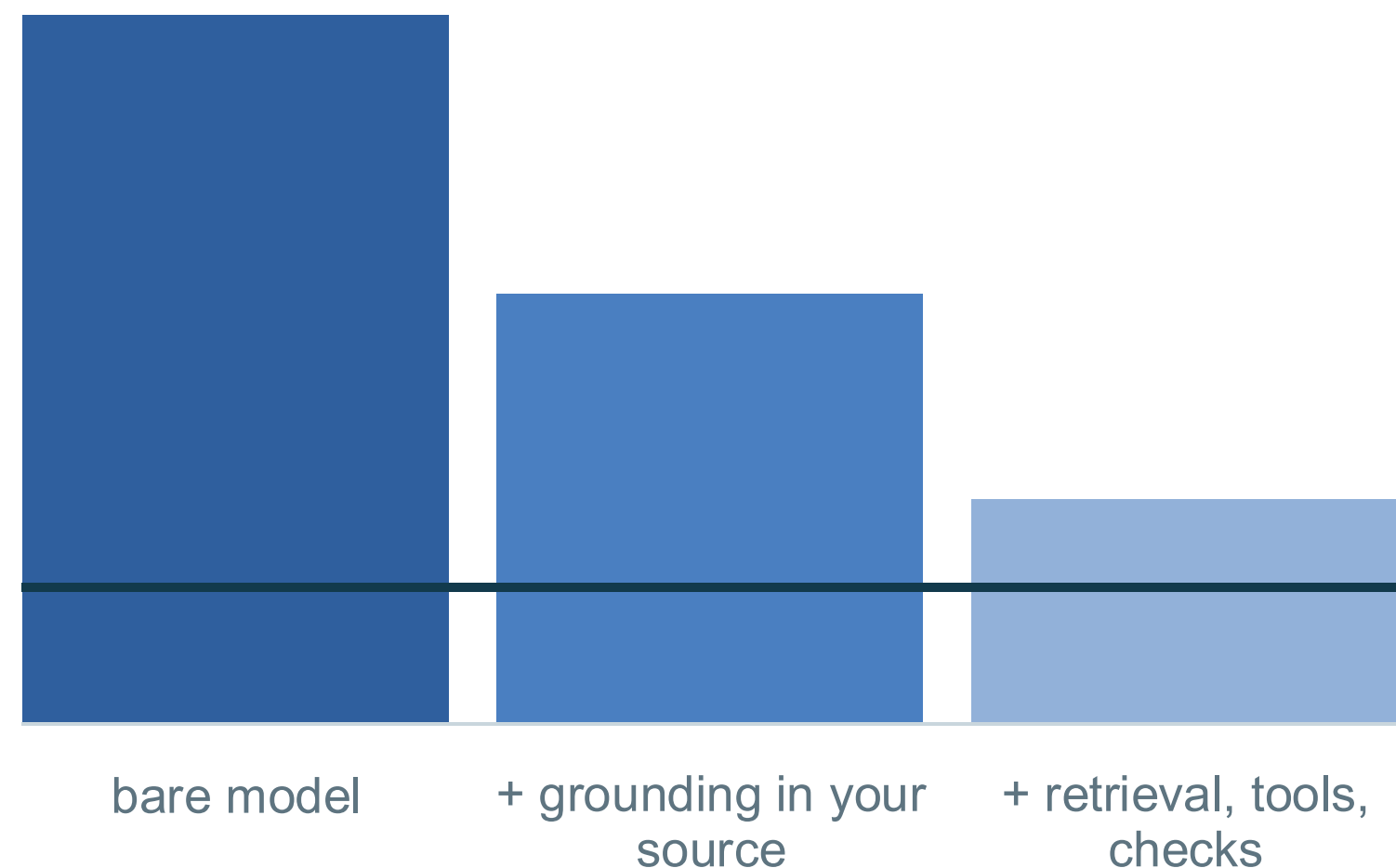
“AI hallucinations can't be stopped — but these techniques can limit their damage,”
Nature news feature, 21 January 2025

Harnessing

RAG, agents, tool use, verification loops:
all of it lowers the rate.

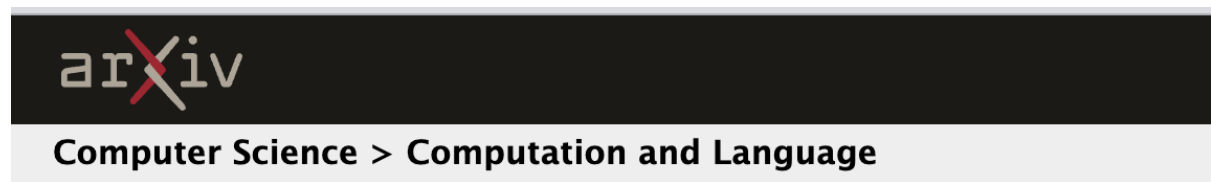
But it works **downstream** of the objective. The model is still a next-token predictor with no notion of truth.

So a fluent, confident, wrong answer stays possible no matter how much you stack on top. That's *Kalai et al.*'s error floor.



— The floor does not reach zero

The bars move with every model release.



[Submitted on 4 Sep 2025]

Why Language Models Hallucinate

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang

Detection by consistency

SelfCheckGPT: sample the same prompt several times and measure whether the answers agree.

RUN 1

Lancet 394, 1712–1721

RUN 2

NEJM 381, 2103–2114

RUN 3

Diabetes Care 42, 731–740

Scored by

- BERTScore
- QA
- n-gram overlap
- NL inference
- prompting the model itself.



Cost: $n \times$ inference, useful for black-box models

DO IT BY HAND

Ask it in three fresh chats and compare by eye. No scoring needed. When it knows the answer, the repeats agree; when it's making something up, they usually don't.

Uncertainty signals

Where the API exposes them:

- **log probabilities**
- hidden-state and attention scores
- embedding similarity between the answer and its source

Semantic entropy clusters repeated samples by *meaning* rather than wording, and flags the ones that scatter.

nature

Explore content

About the journal

Publish with us

nature

articles

article

Article

Open access

Published: 19 June 2024

Detecting hallucinations in large language models using semantic entropy

Sebastian Farquhar

Jannik Kossen

Lorenz Kuhn

Yarin Gal

Nature

630

625–630 (2024)

Cite this article

Farquhar, Kossen, Kuhn & Gal (2024), *Nature* 630, 625–630

Token probability tells you how confident the model is in its *wording*, not in whether the fact is true.

Lindqvist et al. (2019), Lancet 394, 1712–1721

0.94

Token-level confidence

Looks fine. The citation does not exist. High confidence

Correct *format*, highly predictable.

Consistency check: same question, sampled 4 times

A

Lindqvist et al. (2019), Lancet 394

B

Karlsson et al. (2018), BMJ 362

C

Andersson et al. (2020), JAMA 323

D

Nilsson et al. (2017), NEJM 377

4 samples, 4 different meanings: high entropy.

<https://www.youtube.com/watch?v=eiL6egIX4ms>

Grounding helps

RAG grounds the answer in retrieved passages. It **moves** the failure surface: “does the model know this?” becomes “was the right passage retrieved, and did the model use it?” Zhang & Zhang (2025), Mathematics 13(5), 856



RETRIEVAL FAILURE · 4 CAUSES

The passage never reached the model.

Data sources · queries · retrievers · retrieval strategy

GENERATION DEFICIENCY · 5 CAUSES

It was there and the model ignored it.

Context noise · context conflict · middle curse · alignment · capability boundary

Feeding the model your data doesn't erase what it already believes. It just puts the two in competition.

FINE-TUNING ON NEW FACTS

Teaching it new knowledge raises the rate.

New-knowledge examples are learnt slowly, and hallucination rises as they are. Gekhman et al. (2024)

RETRIEVAL IT DOESN'T READ

It prefers its memory over your evidence.

RAG models favour faulty passages supplied over internal memory. Tug-of-War (2024)

RagFlow on UCloud

Corpus is yours: your protocols, your papers, your SOPs,
not the open internet. Data stays inside UCloud.

Every answer comes back with the **chunks it was built from**, shown by default. If there's no chunk behind a claim, the model wrote it.



RAGFlow ☆

RAGFlow is an open-source RAG (Retrieval-Augmented Generation) engine based on deep document understanding.

Job information

Flavor

Default

Version

0.27.1

Job name

Example: Run with parameters XYZ

Hours

1

+1

+8

+24

Service provider

DeiC Interactive HPC (SDU/K8s)

Machine type

No machine type selected

Import

Storage

Folder #1

No directory selected

Connectivity

Public link #1

Public links make your job accessible through a web browser. Anyone with the link can access your application.

No public link selected

Private network #1

Connect this job to a network of other jobs.

No private network selected

Parameters

Data volume: DATA_DIR

Import directory for data persistence. If empty, a new directory tree is generated.

No directory selected

Job report sample rate

Sets sampling rate for resource utilization for the job report (default is 250 ms)

250 ms

Initialization

Shell script to run during startup. File format: Bash script (.sh).

No file selected

In summary

- 1 Source constrains and quote extraction: ask AI to find sources and anchor statements to direct quotation**
- 2 Uncertainty permission: tell the AI it's okay if it doesn't know**
- 3 Start a new chat and ask AI to find errors**
- 4 Cross-reference with verified sources and be sceptical**
- 5 Double-check facts and ask follow-up questions**
- 6 COT Reasoning: step-by-step reasoning so uncertainty becomes visible**

4

Critical evaluation

What to do at your desk tomorrow

A working checklist

- 1 Ask what you could verify.** Never ask for a fact you have no way to check.
- 2 Scale scrutiny to consequence.** A brainstorm and a methods section are not the same risk.
- 3 Check every citation at source.** DOI or it didn't happen, including the ones that look right.
- 4 Rerun in a fresh chat.** Divergence between runs marks the fabricated part.
- 5 Give it the source.** Grounded in your document beats recalled from training.
- 6 Log model, version and date.** The same prompt will not reproduce next month.

Specificity with no source behind it is the giveaway: exact page numbers, precise percentages, named authors you cannot find.

Use an LLM to work on what you give it, not as a source of facts.

- 1 Every model hallucinates. The rate differs; the category does not.
- 2 Confidence is trained, not earned. Fluency says nothing about correctness.
- 3 Harnessing moves the failure somewhere you can inspect. That is the win.
- 4 The verification burden stays with whoever signs the paper.

Resources

AA-Omniscience: knowledge & hallucination leaderboard

artificialanalysis.ai/evaluations/omniscience

Hallucination Mitigation for Retrieval-Augmented LLMs: A Review, Zhang & Zhang 2025, Mathematics 13(5), 856

doi.org/10.3390/math13050856

Why Language Models Hallucinate, Kalai et al. 2025

arxiv.org/abs/2509.04664

Semantic entropy, Farquhar et al. 2024, Nature 630, 625–630

doi.org/10.1038/s41586-024-07421-0

Does fine-tuning encourage hallucination? Gekhman et al. 2024, EMNLP

arxiv.org/abs/2405.05904

Lost in the Middle, Liu et al. 2024, TACL 12, 157–173

arxiv.org/abs/2307.03172

AI Hallucination Cases Database, D. Charlotin

damiencharlotin.com/hallucinations

The Car Wash Test: 53 models, Opper.ai

opper.ai/blog/car-wash-test

SelfCheckGPT, Manakul et al. 2023, EMNLP

arxiv.org/abs/2303.08896

Fabricated references in medical content, Bhattacharyya et al. 2023, Cureus 15(5):e39238

doi.org/10.7759/cureus.39238

Survey on Hallucination in LLMs, Huang et al.

arxiv.org/abs/2311.05232

AI hallucinations can't be stopped, Nature news feature, Jan 2025

nature.com/articles/d41586-025-00068-5