Hallucination, verification, and trust

Why models make things up — and how to build trust around that.

2023: Six federal cases that never existed.

A US attorney filed a brief citing them. They had names. Docket numbers. Holdings. Everything — except being real.

Varghese v. China Southern Airlines, 925 F.3d 1339 (11th Cir. 2019)

Martinez v. Delta Air Lines, 2019 U.S. Dist. LEXIS 89156 (S.D. Fla.)

Zicherman v. Korean Air Lines Co., 516 F.3d 1237 (D.C. Cir. 2008)

Plausible. Formatted. Fluent. Fabricated.

The model wasn't lying. It was doing what it was trained to do.

THE MECHANISM

Hallucination is not a bug.

It's the same mechanism that makes LLMs useful, pointed in a direction you didn't want.

LIKELY & TRUE

"The capital of France is"

→ **"Paris."**

LIKELY & FALSE

"Citations supporting this argument include"

→ **Plausible citations to made-up cases.**

From inside the model, these two operations feel identical.

Two practical consequences.

CONFIDENCE ≠ CORRECTNESS

The fluency is constant.
The accuracy is variable.

There's no tell in the output itself. Humans read fluency as expertise. That heuristic is broken for LLMs.

THIS WON'T GET FIXED

Better models hallucinate less.
Not zero.

The mechanism producing occasional bad outputs is the same one producing all the good ones. You don't get the second without some risk of the first.

THE REFRAME

~~*Stop asking: "Is this true?"*~~

A yes/no question the model can't answer and you can't check at output time.

Start asking: "What verification does this task need?"

A question that's answerable, actionable, and different for every task.

Design around hallucination. Don't wait for it to disappear.

The verification ladder.

Four rungs. Each with a matching verification posture.

RUNG 04	**Regulatory · legal · medical · safety**	*Trust only with tools of record*
RUNG 03	**Meaningful stakes**	*Trust after human review*
RUNG 02	**Low stakes, recoverable**	*Trust with sanity check*
RUNG 01	**Zero stakes**	*Trust freely, catch later*

RUNG 01

Zero stakes.

The cost of wrong is the time to notice and redo.

TYPICAL TASKS

- Brainstorming
- First-draft marketing copy
- Exploring an idea
- Personal note-taking

VERIFICATION POSTURE

Trust freely.

Catch errors when you notice them. Don't build a verification process for the cost-free.

RUNG 02

Low stakes, recoverable.

Cost of wrong: modest rework.

TYPICAL TASKS

- Summaries for personal use
- Internal notes and meeting recaps
- Rough analysis informing a decision
- Research preliminary to actual work

VERIFICATION POSTURE

Trust with sanity check.

Skim for obvious errors. Ground anything that'll be repeated or referenced.

RUNG 03

Meaningful stakes.

Cost of wrong: embarrassment, eroded trust, possibly money.

TYPICAL TASKS

- Content that will be published
- Analysis someone will act on
- Recommendations to clients
- Public-facing communication

VERIFICATION POSTURE

Trust after human review.

Every claim checked. Citations followed. Facts spot-tested. The human is the last line before output becomes real.

RUNG 04

Regulatory, legal, medical, safety.

Cost of wrong: legal exposure, physical harm, regulatory penalty.

TYPICAL TASKS

- Court filings
- Medical info a patient will act on
- Financial advice
- Compliance-sensitive content

VERIFICATION POSTURE

Trust only with tools of record.

AI drafts. Authoritative systems (case databases, clinical guidelines, certified calculators) verify every factual claim.

How to place a task on the ladder.

Two questions. Answer honestly, and the rung is usually obvious.

THE CONSEQUENCE TEST

If the output is wrong and goes out, what happens?

"I'll redo it."	→ *Rung 1 or 2*
"Someone will act on this."	→ *Rung 3*
"Legal, regulatory, or physical harm."	→ *Rung 4*

THE AUDIENCE TEST

Who sees this output?

Only you	→ *Rung 1 or 2*
A client	→ *Rung 3*
A court, regulator, or patient	→ *Rung 4*

Grounding reduces. Never eliminates.

RAG, tool use, citations — probability improvements, not correctness guarantees.

RAG system	*Misquotes a source it correctly retrieved.*
Model with calculator	*Types the wrong number into the calculator.*
Model asked to cite	*Hallucinates the citation, then confabulates a plausible quote from it.*

"Add RAG, therefore no more hallucinations" re-creates the problem one layer deeper.

What this changes.

Two shifts once you accept the verification ladder.

TRUST GETS GRANULAR

Not "do I trust AI?" but "do I trust this output, for this task, at this rung, with this verification?" Different answer every combination. Operating at the right resolution.

ADOPTION GETS UNBLOCKED

Much of "we can't use AI because it hallucinates" is really "we're treating every task as rung 4." Specify the rung. Apply matching verification. Unlock the rung 1 and 2 uses.

Takeaways

01

Hallucination is a feature, not a bug.

The same mechanism that produces good outputs produces occasional bad ones. Design around it.

02

Four rungs. Match verification to stakes.

Zero / low / meaningful / regulatory. Wrong posture for the rung is the real failure mode.

03

Grounding reduces risk. It doesn't remove it.

RAG, tools, citations — all probability improvements, not correctness guarantees. Verify anyway.