The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Why language models hallucinate — and why the way we grade them keeps it that way

The confident falsehoods we call "hallucinations" are not a mysterious glitch: some are a statistical by-product of training, and they persist because mainstream benchmarks reward a confident guess over an honest "I don't know." A case study on four frontier models shows that stating the scoring rules in the prompt ("open rubrics") reverses that incentive.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

Why models guess, and why we taught them to

Ask a large language model for a stranger's birthday and it may answer "March 7" with the steady confidence of someone reading it off a card — and be wrong, three times running, with three different dates. The authors give exactly this kind of example: leading models asked a plain factual question — a person's birthday, or what an obscure acronym stands for — each confidently inventing a different answer, none of them correct. The industry's word for this is *hallucination*, which makes it sound like a fault in perception. The paper's first move is to take the mystery out of it.

Start with how a model is built. In its first and largest training stage it learns, in effect, what fluent language looks like by reading an enormous amount of text. Now take a fact with no pattern behind it — one particular person's birthday. If that date appeared in the training text once, or never, there is nothing for a pattern-learner to grab onto: the answer is, from the model's point of view, arbitrary. The authors make this precise by borrowing an old idea (Alan Turing's, from a different problem): if one in five birthdays shows up only once in the data, a model should be expected to get at least one in five of them wrong — not because it is broken, but because there was never anything there to learn. (By the same logic, models almost never get a country's capital wrong: those appear constantly.) They argue, with some care, that telling a true statement from a plausible false one is itself a hard problem, and that producing *only* true statements is at least as hard as that. A floor of error is baked in.

The idea underneath: how to count what you haven't seen yet

This rests on a genuinely clever idea — older than language models, and worth meeting properly.

Start with a bag of coloured balls. You don't know how many colours it holds. You draw 100, one at a time, and tally them: red 40, blue 25, green 15, yellow 5, purple 3, orange 2 — and then **ten different colours that each turn up exactly once.**

Now the question Turing actually faced, on a quite different problem: *what is the chance the next ball is a colour you haven't seen at all?* You cannot count what you have never drawn — but you **can** count the colours you have seen exactly once, the "singletons." The trick, called **Good–Turing estimation**, is that the share of your draws that are singletons estimates the probability still hiding in the colours you have not seen. Ten of your hundred draws were once-only colours, so the chance the next ball is a brand-new colour is about **10 / 100 = 10%.**

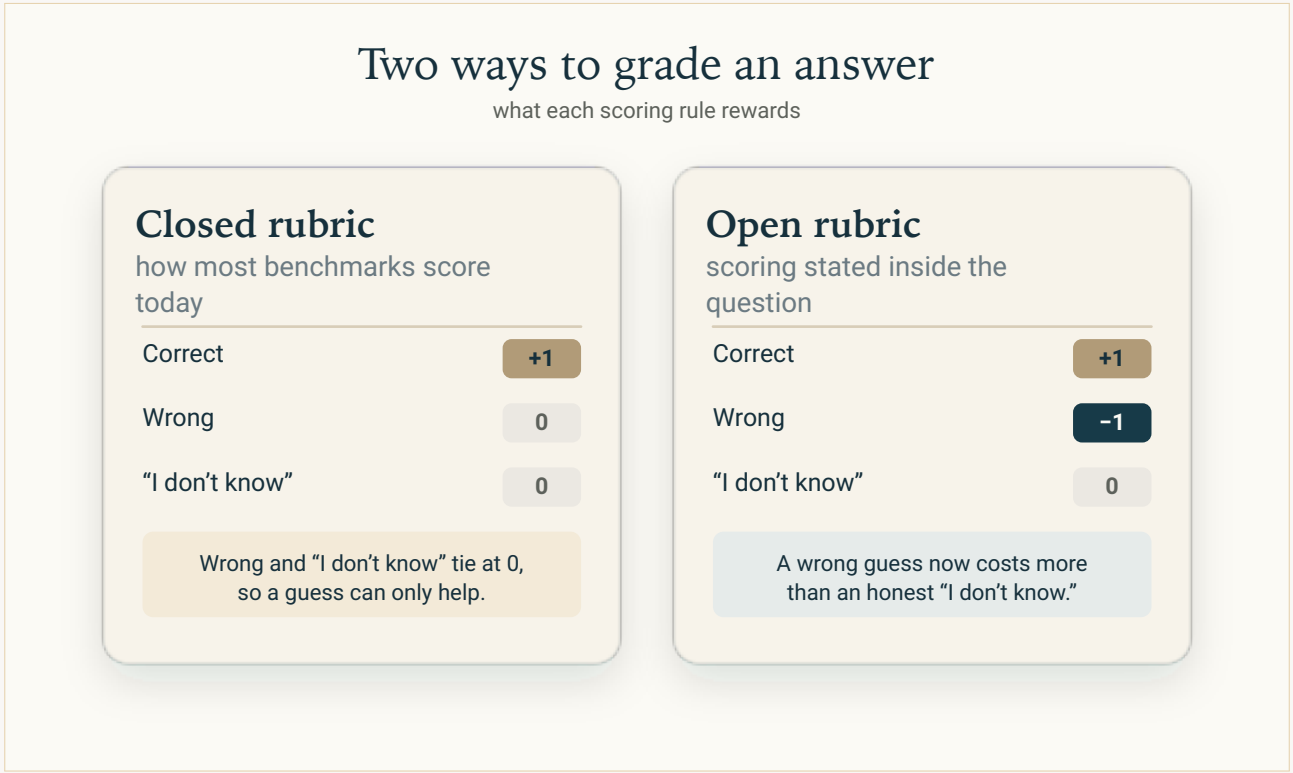
Those once-seen colours are not *mistakes*. They are a measurement of your own ignorance: many colours turning up once is the sample's way of telling you the world holds more that you simply have not drawn yet.

Now swap colours for birthdays, and the bag for the model's training text. Suppose that, among the birthdays it saw, **one in five appears exactly once.** Same trick: about a fifth of the probability lives in birthdays the model has effectively never seen — and a birthday has no pattern to fall back on (you cannot *work out* someone's birthday). So a date seen once, or never, is a coin the model cannot weight, and it will be wrong on roughly **one in five** of them. No amount of cleverness fixes this: there was nothing there to learn.

That is the whole argument in miniature: the singleton rate measures how much of the world is unlearnable *from this data*, and that becomes a **floor** under the errors. It is also why a model

> almost never misses a capital city — *Paris* appears constantly, its singleton rate is near zero, so there is plenty to learn.

That explains where hallucinations come from. It does not explain why they *survive* — why models, after all the later training meant to make them helpful and honest, still bluff rather than admit doubt. Here the paper's analogy is almost uncomfortably apt. Picture a student in an exam who does not know an answer. If a blank scores zero and a guess might score one, the grade-maximising move is to guess — confidently, specifically, never "I'm not sure." Students learn this. So, it turns out, do models — because we grade them the same way. The authors went through the benchmarks the field actually competes on, the leaderboards models are tuned to climb, and found that nearly all of them give "I don't know" exactly the same score as a wrong answer: zero. Under that rule, a model that always guesses will beat an otherwise-identical model that honestly flags its uncertainty. We are, in a fairly literal sense, scoring them into it.



Benchmarks can make guessing rational. Under the scoring most benchmarks use (left), a wrong answer and an honest "I don't know" both score zero — so a guess can only help. State the rules in the question, with a penalty for being wrong (right), and abstaining when unsure can become the better move. It changes what the test rewards; it does not, by itself, solve hallucination.

This is the part worth keeping, because it runs against the usual headline. Hallucination is often sold as an inevitable, almost mystical limit of the technology. The paper disputes that on both counts. The pretraining floor is not a mystery — it is ordinary statistical error, the kind machine learning has understood for decades. And the persistence is not inevitable — it is, in part, an incentive we built

and could change. A system that simply declined to answer when unsure would not hallucinate at all; the reason deployed models don't behave that way is that our scoreboards punish the refusal.

What the authors did

The paper has three parts. First, a mathematical argument that some hallucination is statistically forced during pretraining, by showing that "generate only valid text" is at least as hard as a binary "is this statement valid?" classification problem. Second, an argument — backed by a survey of ten influential benchmarks — that mainstream accuracy-style metrics reward guessing over abstaining. Third, a proposed fix and a case study testing it: **open-rubric** evaluations, where the scoring is stated inside the question itself (for example, "a correct answer scores 1, a wrong one −1, so abstain if you are less than 50% sure"), so that a model can tell when honesty is being rewarded. They try this on four frontier models — Google's Gemini 3 Pro, OpenAI's GPT-5, xAI's Grok 4 and Anthropic's Claude Opus 4.5 — using SimpleQA's 4,326 factual questions. They are explicit that the case study is illustrative, "not a controlled evaluation across models" (default settings, no tuning, no cost normalisation).

What they found

- **Pretraining forces some error.** The rate at which a model emits confident falsehoods is bounded from below by (roughly twice) the error rate of the best "is this statement valid?" classifier built from it. For facts with no learnable pattern, that floor is at least the *singleton rate* — the fraction of facts that appear exactly once in training. Some hallucination is unavoidable even with perfectly clean data.
- **Grading rewards guessing — concretely.** Under ordinary correct/incorrect scoring, never abstaining is the optimal strategy, and the authors' survey finds the vast majority of popular benchmarks score "I don't know" as simply wrong. A vivid example from their own side: on the SimpleQA test, raw accuracy slightly *favours* OpenAI's o4-mini — which answers almost everything and is wrong more than three-quarters of the time — over GPT-5-mini, which makes far fewer mistakes because it abstains when unsure. The more reckless model looks better on the scoreboard.
- **Open rubrics flip the incentive (in their case study).** They test a simple hallucination mitigation (have the model answer twice and abstain if the two answers disagree). Under standard accuracy, the mitigation cuts errors *but also cuts accuracy* — so the metric discourages adopting it. Under open rubrics, the same mitigation comes out ahead for all four models across a range of penalties; and GPT-5-mini — which raw accuracy had penalised for abstaining when unsure — comes out ahead of o4-mini once the scoring is stated openly (n = 4,326 questions per model).

What this probably means

Cutting hallucination is mostly not a matter of inventing more hallucination-specific tests. It is a matter of changing how the mainstream benchmarks score uncertainty, so that admitting "I don't know" is no longer punished. Until the scoreboard changes, reducing hallucination will keep costing models accuracy points and so keep being discouraged — which is why the authors frame the problem as "socio-technical": part better metric, part getting the influential leaderboards to adopt it.

What this does *not* prove

- It does **not** show open rubrics fix hallucination in the wild. The supporting experiment is a small, deliberately **uncontrolled** case study — four models at default settings, one chosen mitigation, one factual-QA test — meant to demonstrate the *incentive flip*, not to rank models or prove general efficacy.
- It does **not** claim grading is the *only* cause. Errors in the training data, genuinely hard problems, and unfamiliar prompts remain separate sources.
- It does **not** support the popular line that hallucinations are inevitable. The authors argue the opposite: a system that answered only checkable questions and otherwise said "I don't know" would never hallucinate.
- It does **not** make the pretraining floor disappear — it explains and bounds it, and the bound concerns confident factual errors, not all model behaviour.
- It does **not** show that open rubrics are *sufficient* on their own. They change what an evaluation rewards; they are not a substitute for retrieval, tool use, or better-calibrated models.

How strong is the evidence?

- The core is **mathematics** — formal lower bounds, not measurements. As a theoretical argument it is sound on its own terms.
- It rests on **deliberately simplified models** of the problem; the authors themselves flag the "false trichotomy" of treating every response as correct, incorrect, or "I don't know," and the idealised "arbitrary facts" setting used for the cleanest bound.
- The benchmark survey is a **small, curated sample** — ten influential evaluations, not an exhaustive audit.
- The case study is **real but limited**: four frontier models, a single mitigation, SimpleQA only, default settings, explicitly "not a controlled evaluation." It is a proof of concept for the incentive argument, not a benchmark result.
- Worth naming the vantage point: three of the four authors are or were employed at **OpenAI**, and the paper argues the field should change how it evaluates models. That is a well-argued position from an interested party, not a neutral outside review — to weigh, not to dismiss. (To its credit, the paper points the critique at its own models, o4-mini and GPT-5-mini, as readily as at others.)

Why it matters

It reframes a heavily hyped problem. “Hallucination” tends to be sold either as a spooky defect or as an immovable wall; this paper makes it ordinary and partly self-inflicted — a statistical floor we can actually understand, sitting on top of an incentive we chose. The wider lesson is quieter and more useful: further progress on reliability may depend as much on *what we measure* as on what we build.

Clean summary

Confident false answers from language models come from two places. The first is statistical: when a fact has no pattern to learn, a model trained to imitate language will sometimes get it wrong, and that floor can be estimated (for instance, from how many facts appear only once in training). The second is incentives: almost every benchmark that models are ranked on scores “I don’t know” the same as a wrong answer, so guessing always wins — to the point that a model wrong three-quarters of the time can outscore a more honest one that abstains. The authors’ proposal is not another hallucination test but “open rubrics”: state the scoring inside the question. In a case study on four frontier models, that flips the incentive so a hallucination-reducing method is rewarded rather than penalised. It is a theory-plus-survey paper with a small, explicitly uncontrolled experiment, peer-reviewed in *Nature*; the fix is promising but not yet shown to work at scale, and hallucinations are argued to be neither mysterious nor strictly inevitable.

No-BS check

What the paper shows: A mathematical lower bound under which some hallucination is forced during pretraining (at least the “singleton rate” for pattern-less facts); a survey finding most leading benchmarks give “I don’t know” no credit; and a four-model case study in which stating the scoring in the prompt (“open rubrics”) makes a hallucination-reducing method win, where plain accuracy had penalised it.

What is plausible but not proven: That open rubrics, rolled into mainstream benchmarks, would meaningfully reduce hallucination in deployed models. The supporting experiment is small and explicitly uncontrolled.

What it does not show: That hallucinations are inevitable (it argues the reverse); that grading is the sole cause; that hallucination can be eliminated outright; that the case study ranks the four models against each other.

Main limitations: A deliberately simplified correct/incorrect/“I don’t know” model (the authors call it a “false trichotomy”); a small curated benchmark survey (ten evaluations); an uncontrolled case study (four models, one mitigation, one test, default settings); and an OpenAI-led argument about how the field should evaluate models.

How much confidence should a general reader have? High that hallucination is neither mysterious nor strictly inevitable, and that mainstream benchmarks currently reward guessing. Moderate that the proposed fix helps — it now has a real proof-of-concept, but not yet a demonstration that it works broadly and at scale.

Sources

Nature 653, 1047–1050 (2026)
https://doi.org/10.1038/s41586-026-10549-w

Why Language Models Hallucinate (arXiv 2509.04664)
https://arxiv.org/abs/2509.04664

hallucinations-paper-experiments (OpenAI)
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals