

The Faithful Liar

Michael P. Laing

`mllaing@post.harvard.edu`

ORCID 0009-0004-5593-5759

Written with Anthropic Claude Fable 5.1

August 2026; revised September 28, 2026

Companion essay to “Which Hallucinations Are Artifacts?” (September 2026) and to “Measure-Relational Algebra: The Core” (September 2026; doi:10.5281/zenodo.22999911). Canonical reference: realql.org.

Two ways to be wrong

A witness takes the stand and describes a meeting that never happened. There are two very different ways this can occur.

The first witness is inventing. Under the pressure of the question, some mixture of suggestion, vanity, and noise assembles a memory on the spot. Ask again tomorrow and the details drift. Ask his brother, who was supposedly also there, and you get a different meeting entirely. The invention is his; it lives in him; it dies with him.

The second witness is reading. Years ago, a clerk copying the minutes made an error — not a random one, but the kind a particular copying procedure reliably makes — and every archive in the city holds a copy descended from that clerk’s page. This witness has a superb memory and consulted the archives diligently. He reports the meeting confidently, in detail, and identically every time you ask. So does his brother. So does the rival family’s archivist, who despises them both but reads the same descended copies.

Both witnesses say a false thing. Only the first is doing anything wrong. The second is telling the truth — about the record. And this difference, invisible in any single answer, is the most consequential fact about him, because everything you might do to fix the first witness will fail on the second. Coach him, cross-examine him, make him more careful, give him *more archives to read*: he gets more confident and more consistent, because the archives agree. The error is not in the witness. It is upstream, in the copying.

The current conversation about machine “hallucination” talks almost entirely about the first witness. This essay is about the second.

What the record did

The core paper on realql.org proves a small theorem with a long reach: when you write down an uncertain measurement as if it were an exact number, and something downstream computes

on it in any way more complicated than adding and scaling, the discarded uncertainty does not politely vanish. It curdles into structure — bends, ceilings, saturations — that the world does not have but the record now does.

This is not hypothetical. For roughly forty years, space physicists observed that a key measure of the solar wind’s grip on Earth’s magnetic field appeared to *saturate*: push the storm harder and the response leveled off, as if nature had installed a ceiling. In 2026 it emerged that nature had installed no such thing. An upstream measurement carried uncertainty; the analysis pipeline had treated it as exact; and that single act of tidying, passed through the nonlinear physics downstream, manufactured the ceiling. Restore the uncertainty and the ceiling evaporates — and the estimated impact of an extreme storm roughly doubles. Forty years of trained, skeptical, professional readers looked at a feature of the record and saw a feature of the world.

Now put a machine in the reader’s chair. A language model is, before it is anything else, a diligent student of a record — the largest record ever assembled, and one that asserts numbers, findings, rankings, and facts almost entirely stripped of the uncertainty of the observations behind them. The model’s training has one job: capture what the record says, as exactly as possible. A companion paper on realql.org works out what this means with some care, and the conclusion is blunt. A competent learner trained on a record like this converges to what the record says — including the structure the record manufactured. The ceiling that is not there is in the lesson. The model learns the lesson.

So when such a model confidently reports the ceiling, it is not hallucinating in any useful sense of the word. It is the second witness. It is faithfully reporting an unfaithful record.

Why more data cannot fix it

The instinctive remedies all aim at the witness. More data. Bigger models. Better training. Each of these makes a learner better at its actual job — matching the record — and the manufactured structure is *in* the record. Not in this document or that one, where more documents would dilute it, but in the overall tendency of the whole record, where more documents *confirm* it. A million fresh readings copied through the same collapsing pipeline agree with the first thousand. The student’s error is not a failure to learn; it is the graduation prize.

This is the dividing line that matters, and it is worth saying plainly, because it cuts the fashionable word “hallucination” in two. Some machine errors really are the first witness: noise, one-off inventions, artifacts of a particular model on a particular day. Those, the ordinary remedies address, and the industry’s progress on them is real. But some errors are the second witness, and against those the entire model-side arsenal does not merely fail — it backfires. Every improvement produces a witness more confident, more consistent, and more persuasive about the clerk’s mistake.

The tell

If the two kinds of error looked alike, this would be a philosophical distinction. They do not look alike. They differ in the one way you can measure from outside: *repetition*.

The first witness’s inventions are his own. Train the same model twice with different random starts, and the inventions differ. Ask a rival company’s model, built by different people on different machines, and its inventions differ again. Independent noise is the signature of confabulation, exactly as eight hundred fairgoers guessing an ox’s weight err independently.

The second witness’s error is the record’s, and the record is shared. Same model, retrained: same wrong answer. Rival model, different architecture, different lab: same wrong answer, because it studied substantially the same library. The wrongness travels with the record, not the model.

Here the story stops being a prediction and touches published measurement. In 2025 a Cornell group examined over three hundred and fifty language models and asked a simple question: when two models are both wrong on a test item, how often are they wrong *the same way*? On one widely used evaluation suite the answer was about sixty percent — far beyond coincidence. And the finding that should stop a reader mid-sentence: the larger and *more accurate* the models, the more their errors agreed — even across different architectures and different companies. If machine error were noise, better machines should err more idiosyncratically, not less. But if a shared record fixes a shared lesson, then better students converge on the same mistakes — because the mistakes are in the syllabus. The best students in the class all miss the same question, and they all give the same wrong answer, and that is how you know the answer key is wrong.

There is an honest caveat, and the technical paper carries it: not everything that repeats is the record’s manufactured structure — shared bad test questions repeat too, and the experiments have to control for that. Repetition points the finger upstream; it takes more work to convict the clerk. The paper specifies that work.

Collapse three times

It would be tidy if the record were the only place the tidying happened. It happens three times.

It happens at the *record*, as above: uncertain observations written down as exact numbers, before the model ever sees them. This is where the manufactured structure is born, and the only place it can be prevented.

It happens at the *mouth*. Whatever a model actually holds about a quantity — and internally it holds a spread of possibilities, weighted — the interface obliges it to speak one token at a time, one committed word after another. A model that holds a value only loosely asserts it to three decimal places, because the pipeline offers no register for “roughly.” The doubt it still had is discarded at the moment of speech.

And it happens at the *report card*. The benchmarks that rank models grade answers right or wrong, full stop. Researchers at OpenAI and Georgia Tech pointed out in 2025 what every student already knows about such tests: if a blank scores zero and a wild guess might score one, the rational test-taker guesses. Models are optimized against these tests, so the industry has been systematically rewarding confident guessing over honest doubt — selecting, generation after generation, for the smoothest possible witness. Their proposed fix, regrading the tests to reward calibrated “I don’t know,” is right as far as it goes. Note where it stops: a witness can only say “I don’t know” about what he knows he doesn’t know. Our second witness fails that condition precisely. He *does* know. The archive told him.

Three collapses, one habit: at every stage where doubt could be carried forward — writing, learning, speaking, grading — the pipeline writes down a point and moves on. Each step is locally reasonable. The composition is a machine for laundering uncertainty into confidence.

The claim, plainly

Machine hallucination is not one thing. Part of it is invention, and the ordinary remedies — more data, better training, stricter grading — are the right medicine and are working. But part of it is faithful reporting of a record that manufactured its own falsehoods at the moment someone wrote uncertain measurements down as exact numbers, and that part has three properties the public conversation has not yet absorbed. It cannot be fixed by any improvement to the model, because the model is not broken. It grows more convincing as models grow better, because fidelity to the record is what better means. And it is detectable now, by its fingerprint: the same wrong answer, delivered independently, by rivals.

The fix belongs where the fault is: in what the record keeps. A memory that carries measurements together with their uncertainty — what that requires, exactly, is the business of the papers this essay accompanies — would starve this class of error at its source. Until then, it is worth learning to ask, of any confident machine answer, the question the space physicists finally asked after forty years: is this a fact about the world, or a fact about the filing?

Sources and precision for the claims sketched here — including the geomagnetic saturation result (Nature, 2026), the Cornell measurement of correlated errors (ICML 2025), the OpenAI / Georgia Tech analysis of grading incentives (2025), and the theorems about what collapse manufactures — are given in the technical papers at realql.org.