Why language models *make things up*

Why a system trained to predict text states false things fluently, what the main explanations rest on, and what has been shown to reduce it.

MITWA LAB, MUMBAI

IN PLAIN WORDS

Ask a language model something it has no answer to and many models will give one anyway, in the same assured voice they use when they are right. Researchers call this hallucination. One reason for it is statistical, since a system trained to produce text like the text it has read will sometimes produce false sentences that look like true ones, most of all about facts it met only once. Another lies in how models are tested. On most tests a wrong guess and an honest "I don't know" both score nothing while a guess might score a point, so guessing wins. Systems that look things up made fewer errors than those that did not in the two studies reviewed here, though none of them was free of errors.

A thesis that was never written

Suppose you ask a chat assistant about a statistician who is respected in her field and little known outside it: "What was the title of her doctoral thesis?" In a reply invented here for illustration, the assistant says at once that the thesis was called "Bayesian Methods for Sparse Longitudinal Data" and was examined in 2011. Both details are made up, and nothing in the wording says so.

Kalai et al. (2025) report a real case of this kind. In May 2025 they asked three popular chat models, none of which searched the web, for the dissertation title of one of the paper's authors. They got three different titles, universities and years, and no model had the right title or year.

The field calls this hallucination. Ji et al. (2023), in a survey of text generation, define it as output that is nonsensical or unfaithful to the source the system was given. A chat model answering from memory has no source document, and Kalai et al. (2025) use a simpler definition, a plausible falsehood.

The word has critics. Hicks et al. (2024) object that to hallucinate is to misperceive, whereas a language model perceives nothing and produces a false sentence by the same process as a true one. They prefer the philosopher Harry Frankfurt's term, bullshit, for speech produced without regard for truth. Farquhar et al. (2024) complain that one word covers errors with different causes, and reserve "confabulation" for answers that are wrong and arbitrary, where asking again may give a different answer. We keep the word because the papers use it, and say for each study what it counted.

Errors that follow from imitating text

In the first of its two training stages, a language model is fitted so that the text it generates is statistically like the text it learned from. Kalai and Vempala (2024) asked what that implies for arbitrary facts, their term for facts with no rule behind them, such as who ate what for lunch on a particular day. In their idealised setting every training document is true, each carries at most one fact and the model generates with no question put to it.

The argument rests on calibration in the sense used for weather forecasts, where a forecaster is calibrated if rain falls on about 30% of the days given a 30% chance. A language model is calibrated in their sense if statements it generates with a given probability occur about that often in real text. Real text keeps producing facts that no sample contained, and the

Good-Turing estimate puts the probability of such unseen facts at about the share of observations seen exactly once. A calibrated model must give about that much probability to statements it never saw, and since it cannot tell which of them are true and plausible falsehoods far outnumber truths, most of it lands on falsehoods. Their theorem assumes that no fact missing from the training text is much more probable than the others, and under that condition it puts a calibrated model's rate of hallucination on arbitrary facts at no less than about the proportion of those facts that the training text stated only once, even when the text contains no errors.

Kalai et al. (2025) widen the argument. Generating only valid statements, they show, is at least as hard as judging whether a given statement is valid, and a model's rate of generating errors is at least about twice its error rate on that yes-or-no task, less correction terms that are small for a calibrated model with far more wrong answers available than right ones. Generation errors then share causes familiar from classification, among them no pattern to learn (birthdays), a model that fits the task poorly (counting the letters in a word) and training text that is itself wrong. Their example, a hypothetical that was not measured, is that if 20% of birthday facts appear once in training, a model fresh from that first stage should get at least 20% of birthdays wrong.

Whether real models and real text meet these assumptions is an empirical matter. One relevant measurement is from Kandpal et al. (2023), who counted, for each test question, the documents in a model's training corpus that mention both its subject and its answer. For the largest BLOOM model, trained by the BigScience initiative (a different model from our companion of the same name), accuracy on TriviaQA rose from 25% to above 55% as that count went from ten to ten thousand. In a separate experiment the authors trained another, smaller model with and without the relevant documents, and its accuracy fell on the questions whose documents had been removed. Accuracy does not separate a wrong answer from no answer, so the study supports the premise that rarely stated facts are poorly learned and cannot say how often an invention fills the gap.

A second route needs no gap in the data, because a model that fits its training text well can also repeat false beliefs that the text repeats. Lin et al. (2022) wrote TruthfulQA, 817 questions that some people answer wrongly because of a misconception. The best model they tested (GPT-3-175B, with instructions the authors label "helpful") gave true answers to 58%, counting non-committal replies as true, against 94% for a human participant, and within a model family the larger models were generally less truthful.

None of this shows that fabrication is unavoidable, although that has been argued. Xu et al. (2024) define a hallucination formally as any disagreement between a model and a computable function that gives the correct output, and prove by a diagonal argument that for any computable model some such function exists on which it errs. That is a statement about existence, and the authors acknowledge that it gives little insight of a probabilistic kind. Kalai et al. (2025) reply that a system which answers only from a fixed store of questions and a calculator, and otherwise says it does not know, would make no errors. The statistics explain where fabrications come from in the first stage of training, and leave open why they survive the second, which is meant in part to remove them.

What a guess earns on a test

Kalai et al. (2025) explain that persistence by incentives. Under the right-or-wrong marking that most benchmarks use, "I don't know" scores nothing, a wrong answer scores nothing and a guess has some chance of a point, so declining is never the best move. Of ten popular benchmarks the authors audit, nine grade this way with no credit for declining, and the tenth uses a rubric under which admitting ignorance may score below a response that contains a fabrication. On their account a model optimised for such tests is being optimised to guess.

Their remedy is to state the terms in the question. The model is told to answer only if it is more than t confident, that a mistake costs $t/(1 - t)$ points, that a correct answer earns one point and that "I don't know" earns none. That declining is never the best move under right-or-wrong marking is proved, and the audit is a count of ten benchmarks, but the causal step (that benchmark scoring is what keeps released models guessing) is an argument. The paper contains no experiment in which a model trained or tested under the proposed scoring fabricated less, and the authors write that existing models "may or may not" behave as the scheme intends.

SimpleQA, built by Wei et al. (2024), shows what the scoring rule does to a ranking. It has 4,326 short factual questions, and each answer is graded as correct, incorrect or not attempted. GPT-4o attempted nearly everything and was correct on 38.2% and incorrect on 60.8%. Claude 3.5 Sonnet declined 35.0%, and was correct on 28.9% and incorrect on 36.1%. By accuracy alone the first model leads by about nine points, and it is also the one that

answered six questions in ten wrongly. A question entered the set only if at least one of four OpenAI models had answered it wrongly, so these are not error rates for everyday use.

Our Technical Note 01, "Memory you can trust", carries this argument about scoring from questions about the world to a companion's recall of a person's own life, where the costlier error is recalling something the person never said. We report there that invented memories in Bloom, our companion, fell from 127 of 400 checks to 0 of 400, with a one-sided 95% bound of 0.75%, which by the note's own standard is not yet enough.

What the model can consult when it answers

A model answering without tools cannot look anything up, and what training left in it is thin for rarely stated facts. Whether anything in it marks the difference between an answer it holds and one it is making up is an empirical question.

Several results point to some such signal. Wei et al. (2024) found that models asked for their confidence were more accurate when they stated more, yet consistently claimed more than their accuracy warranted. Farquhar et al. (2024) use disagreement instead, sampling several answers to a question, grouping them by meaning and computing the entropy over the groups. Averaged over 30 combinations of task and model, this score told wrong answers from right ones with an AUROC of 0.790, against 0.691 for entropy over the wording alone (0.5 is uninformative, 1 is perfect). It catches arbitrary errors only, and the authors say it does not help when a model's outputs are systematically wrong.

Other work looks for the signal inside the network. Azaria and Mitchell (2023) trained a small classifier on the internal activations of two models, OPT-6.7b and LLaMA2-7b, as they processed true and false statements. On held-out topics, with test sets half true and half false, it labelled 71% to 83% of statements correctly, depending on the model. Orgad et al. (2025), whose study covered four open-source models and ten datasets, found on three of those datasets cases where such a classifier could pick the correct answer out of thirty sampled responses although the model showed no preference for it when generating.

How far this goes is disputed. Levinstein and Herrmann (2024) rebuilt the classifiers of Azaria and Mitchell (2023) on another model, LLaMA 30b, and tested them on negations of statements from the original data. Five of six scored below chance, and the authors conclude that the classifiers had found some feature that accompanies truth in the training topics, which is weaker than a representation of truth. Orgad et al. (2025) report that their own detectors, trained on one dataset, mostly carried over to another no better than the model's output probabilities did. These results support the claim that a model's uncertainty is partly readable, and stop short of showing that a model knows when it is wrong.

What retrieval and abstention changed when measured

One remedy retrieves documents and hands them to the model with the question. Min et al. (2023) asked three systems for biographies of 183 people, split each into single facts and had annotators check them against Wikipedia. The share supported was 42.5% for InstructGPT, 58.3% for ChatGPT and 71.5% for PerplexityAI, which searches the web before answering, and for all three it fell as the person became more obscure.

Magesh et al. (2025) tested commercial legal research tools built on retrieval, which their makers had advertised as avoiding hallucination or free of it. On 202 preregistered queries, counting a response as hallucinated if it said something incorrect or cited a source that did not support it, Lexis+ AI and Ask Practical Law AI hallucinated on 17% and Westlaw's AI-Assisted Research on 33%, against 43% for GPT-4 with no documents. The queries were designed to be hard, and the authors warn that these are not rates for ordinary legal work.

Declining to answer is the other lever, and the same studies show its cost. Ask Practical Law AI matched the lowest hallucination rate while giving incomplete responses, either refusals or answers without supporting citations, to 62% of queries, and among queries it did answer it was not significantly more trustworthy than GPT-4. On SimpleQA the model that declined a third of the questions also answered fewer correctly. Farquhar et al. (2024) show that abstention can be steered by a measured signal, since refusing the questions with the highest semantic entropy raised accuracy on the rest. Kalai et al. (2025) note that retrieval leaves the incentive as it was, because a model graded right or wrong still does better by guessing when a search returns nothing useful. In both studies the systems that retrieved fabricated less than those that did not, and in neither did the rate reach zero.

What is not yet known

- Whether scoring benchmarks differently would change what models do. Kalai et al. (2025) argue that it may, and none of the papers reviewed here tests it.
- How much of the error in released models the statistical bounds explain. They need the share of facts stated once in a real training corpus, for which the papers reviewed here give only hypothetical figures.
- Whether the internal signal reflects anything general. Classifiers held up on new topics in one study, fell below chance on negations in another and failed to carry across datasets in a third.
- How often people meet a fabrication in ordinary use. SimpleQA kept only questions some model got wrong, TruthfulQA targets misconceptions and the legal queries were written to be hard.
- Whether progress on short questions carries over to long answers, which Wei et al. (2024) leave open. Min et al. (2023) found facts late in a generated biography less often supported than early ones.

References

Azaria, A., and Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In Findings of the Association for Computational Linguistics: EMNLP 2023, 967–976. aclanthology.org/2023.findings-emnlp.68/

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630. doi.org/10.1038/s41586-024-07421-0

Hicks, M. T., Humphries, J., and Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26, 38. doi.org/10.1007/s10676-024-09775-5

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. Read as arXiv:2202.03629. arxiv.org/abs/2202.03629

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). Why language models hallucinate. arXiv:2509.04664. arxiv.org/abs/2509.04664

Kalai, A. T., and Vempala, S. S. (2024). Calibrated language models must hallucinate. In Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC 2024), 160–171. Read as arXiv:2311.14648. arxiv.org/abs/2311.14648

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 15696–15707. proceedings.mlr.press/v202/kandpal23a.html

Levinstein, B. A., and Herrmann, D. A. (2024). Still no lie detector for language models: probing empirical and conceptual roadblocks. Philosophical Studies, 182(7), 1539–1565 (published online 17 February 2024). Read as arXiv:2307.00175. arxiv.org/abs/2307.00175

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3252. aclanthology.org/2022.acl-long.229/

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies, 22(2), 216–242. Read as arXiv:2405.20362, from which the figures quoted here are taken. arxiv.org/abs/2405.20362

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). FActScore: fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12076–12100. aclanthology.org/2023.emnlp-main.741/

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. (2025). LLMs know more than they show: on the intrinsic representation of LLM hallucinations. In International Conference on Learning Representations (ICLR 2025). arxiv.org/abs/2410.02707

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. (2024). Measuring short-form factuality in large language models. arXiv:2411.04368. arxiv.org/abs/2411.04368

Xu, Z., Jain, S., and Kankanhalli, M. (2024). Hallucination is inevitable: an innate limitation of large language models. arXiv:2401.11817. arxiv.org/abs/2401.11817