

On 25 September 2026 a federal judge in Massachusetts sanctioned a lawyer whose briefs cited cases that do not exist and quoted cases for words they do not contain. For one of the briefs he admitted using AI, and told the court he had believed the enterprise version of his software did not hallucinate cases. Three days later a public database of such episodes reached 2,095 court decisions. The failure is easy to mock and easy to misread. A fluent answer is evidence that a system is good at producing fluent answers; whether it is right depends on three things the fluency conceals: whether the task resembles the ones on which the system was tested, whether its confidence carries any information, and whether competence at one task says anything about its neighbour. Weather forecasters, epidemiologists and designers of automated control rooms met each of those problems long before chatbots did.

1. A brief that sounded right

The order, signed by Judge Angel Kelley of the United States District Court for the District of Massachusetts, concerns a lawsuit seeking to recover a judgment from insurers, and runs to 19 pages. It opens by noting that "the legal news outlets regularly report another brief or opinion that was published with fictitious citations or facts", and continues: "This is another one of those cases." The court found that one of the plaintiff's filings "cites at least five other cases for quotations they do not contain, includes a fictitious Westlaw citation, omits authority for at least one citation, and seriously misquotes at least four cases", and that another "also cites fictitious cases".

The explanation offered to the court is the instructive part. According to the order, the lawyer

admitted to using AI to draft the Strike Opposition, explaining that he believed that the enterprise-level version of AI software he was using did not hallucinate cases

He said that brief had been an earlier draft filed in haste, attributed errors in two other filings to a bout of influenza, and told the court he had since introduced citation checks. The court was unmoved on the central point:

There is no rule against the use of AI in researching and drafting legal papers, but it must be utilized responsibly.

"The use of AI does not diminish an attorney's professional and ethical obligations under Rule 11," the order continues, and it was "no excuse" that he had not known AI could generate fake citations. The sanction was the other side's fees and costs, capped at \$10,000, and the revocation of his permission to appear in the

case. The parties must report by 23 October 2026 whether they have agreed the fees; a status conference is set for 9 November.

Three days after the order, on 28 September, Damien Charlotin, a legal researcher, updated his AI Hallucination Cases Database, which then listed 2,095 decisions. The database is careful about what it counts. It covers decisions in which a court or tribunal "explicitly found (or implied)" that a party relied on hallucinated material, plus some where AI use was alleged but not confirmed—"a judgment call on my part", in Mr Charlotin's words—and it states plainly: "It does not track the (necessarily wider) universe of all fake citations or use of AI in court filings."

Counted by decision date from the database's published file, the entries grow slowly and then very fast. The file contains 77 decisions dated through 2024, 932 through 2025 and 1,798 through June 2026. From October 2025 through August 2026 the monthly figure moved between roughly 100 and 180. The database's own filters list 1,208 entries involving people representing themselves, 829 involving lawyers and 33 involving judges (some entries involve more than one).

Figure 1. Court decisions addressing AI-hallucinated material, by month of decision, January 2024 to August 2026



Source: Damien Charlotin, *AI Hallucination Cases Database*, CSV download, last updated 28 September 2026, read 30 September 2026; monthly counts tallied from the file's decision dates. The database counts court decisions that address hallucinated material, not fake citations filed, and has no count of filings to divide by, so the line is not an error rate. September 2026 (61 decisions to the 25th) is omitted as incomplete, and recent months may rise as late-reported decisions are added.

▼ Table view

Figure 1. Court decisions addressing AI-hallucinated material, by month of decision, January 2024 to August 2026

Month of decision	Decisions in the database
Jan 2024	2
Feb 2024	4
Mar 2024	4
Apr 2024	3
May 2024	2
Jun 2024	2
Jul 2024	6
Aug 2024	8
Sep 2024	5
Oct 2024	5
Nov 2024	10
Dec 2024	10
Jan 2025	16
Feb 2025	16
Mar 2025	25
Apr 2025	30

Month of decision	Decisions in the database
May 2025	46
Jun 2025	47
Jul 2025	82
Aug 2025	87
Sep 2025	100
Oct 2025	122
Nov 2025	130
Dec 2025	154
Jan 2026	138
Feb 2026	133
Mar 2026	182
Apr 2026	128
May 2026	154
Jun 2026	131
Jul 2026	128
Aug 2026	108

The line is easy to over-read. The count could rise with greater use of these tools, with closer checking of citations by judges and opposing lawyers, or with fuller reporting of decisions to the database, which describes itself as "a work in progress"; it measures none of these. Nothing in it says how often any particular tool invents a case.

2. Two readings that mislead

2.1 "It passed the exam, so it can do the work"

In March 2023 OpenAI's technical report for GPT-4 stated that the model passed "a simulated bar exam with a score around the top 10% of test takers". A reader could be forgiven for concluding that such a system can be trusted with legal research. The lawyer in Massachusetts reached a version of the same conclusion about a premium product.

The best test of that conclusion predates the order by more than two years. In a paper posted in May 2024, researchers at Stanford and Yale—Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher Manning and Daniel Ho—reported a pre-registered evaluation, run between March and May that year, of the leading commercial legal research tools. Their providers had described retrieval-augmented generation, in which the system looks up real case law before writing, as "eliminating" (Casetext) or "avoid[ing]" (Thomson Reuters) hallucinations, or as guaranteeing "hallucination-free" citations (LexisNexis). The

researchers reported that hallucinations were "reduced relative to general-purpose chatbots (GPT-4)", and that the tools from LexisNexis and Thomson Reuters "each hallucinate between 17% and 33% of the time" on their questions.

Claim or result	Source	Date
GPT-4: simulated bar exam "around the top 10% of test takers"	OpenAI, GPT-4 Technical Report	March 2023
Retrieval "eliminating" or "avoid[ing]" hallucinations; "hallucination-free" citations	vendors, as quoted by Magesh et al.	2023
Lexis+ AI, Westlaw AI-Assisted Research, Ask Practical Law AI: 17%–33% hallucination	Magesh et al., arXiv 2405.20362	30 May 2024
Lawyer "believed" the enterprise version "did not hallucinate cases"	U.S. District Court, D. Mass., sanctions order	25 September 2026

The test is more than two years old, and its lesson is not tied to the versions tested: a score earned on one set of questions, or a label printed on a product, describes the conditions under which it was earned, and does not travel automatically to a new task.

2.2 "So the tools are useless"

The opposite reading is as wrong. A count of court decisions is not a failure rate, and the most careful studies of these systems find large gains on some tasks alongside losses on others. The study described in section 3.3, of 758 consultants at Boston Consulting Group, found that those using GPT-4 completed more tasks, faster and at higher quality on most of its assignments, and did worse on one. Blanket distrust fits that evidence no better than blanket trust. The useful question is narrower: what evidence supports this answer, in this setting, and how would a mistake be caught?

3. The idea: three ways a fluent answer fails

3.1 Distribution shift: the flu detector that learned winter

In February 2009 a team of Google researchers led by Jeremy Ginsberg reported in *Nature* that queries typed into Google's search engine could estimate the level of influenza-like illness in America "consistently 1-2 weeks ahead of CDC ILI surveillance reports". Four years later the system's estimates had drifted far from the official ones. In the words of a 2014 analysis in *Science* by David Lazer and colleagues, "Nature reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC)". The system, they found, "has missed high for 100 out of 108 weeks starting with August 2011".

The authors proposed two explanations. The first concerned how the model had been built: it had picked up terms that tracked the winter season rather than the illness, and it "completely missed the nonseasonal 2009 influenza A–H1N1 pandemic".

In short, the initial version of GFT was part flu detector, part winter detector.

The second reason, which they called a "more likely culprit" for the later errors, was what they termed algorithm dynamics: "the changes made by engineers to improve the commercial service and by consumers in using that service". Google's search engine and its users kept changing underneath the model.

That second mechanism is the core of **distribution shift**: a system is built and checked on one slice of the world and then used on another, and its measured accuracy describes the first slice only. The model need not change for its performance to change. Statisticians distinguish several varieties—the mix of inputs can change (covariate shift), the frequency of the outcomes can change (label or prior shift), or the relationship between input and outcome can itself move (concept drift)—but the practical point is common to all of them. A change in conditions does not always degrade a model; it removes the grounds for assuming it has not.

A medical example shows the same shape without the complications of a search engine. In November 2018 John Zech and colleagues published in *PLOS Medicine* a study that trained and evaluated pneumonia-detecting networks using 158,323 chest radiographs from three hospital systems. A model trained on two of them scored an area under the curve (AUC, a measure of discrimination, not a percentage of correct answers) of 0.931 on test images from those hospitals, and 0.815 at a third. A separate network, trained to identify hospital systems, named the source correctly for 99.95% of test radiographs from the National Institutes of Health and 99.98% from Mount Sinai. The authors' conclusion is modest and exact: "Estimates of CNN performance based on test data from hospital systems used for model training may overstate their likely real-world performance."

A more famous story—of a military network that learned to detect sunny weather instead of tanks—is best left untold as fact. Gwern Branwen, who traced its many variants, concludes that "it is definitely not real as usually told". The pneumonia study is the documented version of the same failure.

For a language model, the exam question and the live brief are different slices of the world. The exam arrives complete, with its facts supplied; the brief needs cases that exist, in the right court, saying what the writer claims.

3.2 Calibration: what the weather bureau knew

A forecaster is **calibrated** if, across all the days on which rain was given a 70% chance, it rained on about 70% of them. Calibration is not accuracy. A calibrated forecaster who says 70% still sees dry weather on three such days in ten; a forecaster who always states the long-run average rainfall can be calibrated and useless.

The problem of scoring such forecasts was set out in January 1950 by Glenn Brier of the U.S. Weather Bureau, in the *Monthly Weather Review*. The difficulty, he wrote, was that a forecaster might choose "to let it do the forecasting for him by 'hedging' or 'playing the system'":

This may lead the forecaster to forecast something other than what he thinks will occur

Brier proposed a scheme "that cannot influence the forecaster in any undesirable way", and argued that this "is the case when forecasts are expressed in terms of probability statements". The score that now carries his name measures the overall quality of probability forecasts, of which calibration is one part; it is not a pure calibration measure. The principle that mattered was the incentive: a scoring rule should reward the forecaster for saying what he actually believes.

Language models make that principle newly urgent. Underneath its prose, a model assigns probabilities to the words it might produce next. OpenAI's GPT-4 report measured whether those probabilities meant anything. On a subset of the MMLU multiple-choice test, it reported that "the pre-trained model is highly calibrated (its predicted confidence in an answer generally matches the probability of being correct)", and that "after the post-training process, the calibration is reduced". The caption of its Figure 8 is blunter: "The post-training hurts calibration significantly."

Figure 2. GPT-4's calibration error on an MMLU subset, before and after post-training



Source: OpenAI, GPT-4 Technical Report, arXiv 2303.08774, March 2023, Figure 8 (values as printed on the plot). Confidence is the probability the model assigned to each of the A/B/C/D answer letters, not a confidence the model stated in words. OpenAI's own experiment on private checkpoints; no independent reproduction was found.

▼ Table view

Figure 2. GPT-4's calibration error on an MMLU subset, before and after post-training

GPT-4 checkpoint	Expected calibration error (lower is better)
Pre-trained model	0.0
Post-trained model (PPO)	0.1

Two limits travel with that result. The measurement is of answer-letter probabilities on one multiple-choice test, not of the certainty conveyed by a paragraph of prose; and it is OpenAI's own comparison of checkpoints nobody else can inspect. The broader point does not depend on either. In an ordinary answer none of those probabilities is shown to the reader. A sentence built around an invented case can read exactly like a sentence built around a real one. Sounding certain is a property of the writing, not a calibrated confidence scale.

3.3 Jagged capability

The third idea was named by Andrej Karpathy, an AI researcher, in a post on X on 25 July 2024. "Jagged Intelligence", he wrote, was his word for the fact that state-of-the-art models "can both perform extremely impressive tasks (e.g. solve complex math problems) while simultaneously struggle with some very dumb problems". His examples included judging whether 9.11 or 9.9 is the larger number. The summary sentence is the one worth keeping:

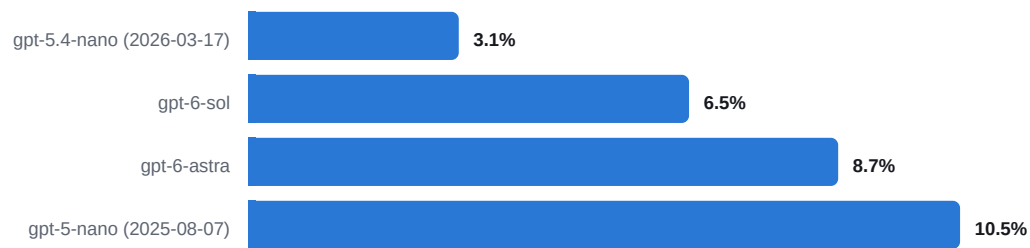
Some things work extremely well (by human standards) while some things fail catastrophically (again by human standards), and it's not always obvious which is which

He drew the contrast with people, "where a lot of knowledge and problem solving capabilities are all highly correlated". A colleague who can write a strong legal argument can usually also check that a case exists. In these systems the correlation is weaker, so success on a hard task says less than expected about an easy one beside it.

An early measurement of the phenomenon came ten months earlier, in a working paper by Fabrizio Dell'Acqua and eight co-authors from Harvard Business School, Wharton, Warwick, MIT and Boston Consulting Group, dated 22 September 2023. It enrolled 758 BCG consultants, "about 7% of the individual contributor-level consultants", and gave them realistic consulting tasks with or without GPT-4. The authors' premise was that "some tasks are easily done by AI, while others, though seemingly similar in difficulty level, are outside the current capability of AI". On 18 tasks inside that frontier, consultants using AI "completed 12.2% more tasks on average, and completed tasks 25.1% more quickly", with results judged more than 40% higher in quality. On one task selected to fall outside it, they were "19 percentage points less likely to produce correct solutions compared to those without AI": the control group was right about 84.5% of the time, and the two AI groups 60% and 70%.

The outside task was chosen by the researchers to catch the model out, so the 19-point loss is not a rate for consulting work in general. What the study shows is that tasks which looked alike fell on different sides of the boundary.

The same unevenness appears between models, and between tests of a single model. Artificial Analysis, an independent evaluator, reported on 9 September 2026 that OpenAI's GPT-6 Astra cut its hallucination rate on the firm's AA-Omniscience knowledge test from 92% for its predecessor, GPT-5.6 Sol, to 51% at maximum effort. The rate is an unusual one: the share of wrong answers among all responses that were not fully correct, including partial answers and abstentions. Vectara's leaderboard, which measures something different—whether a model's summary of a supplied document adds unsupported facts—was updated on 22 September and ranks GPT-6 Astra behind gpt-6-sol and the much smaller gpt-5.4-nano.

Figure 3. Share of document summaries containing unsupported content, selected OpenAI models

Source: Vectara, hallucination leaderboard (GitHub README), last updated 22 September 2026. Each model summarises the same documents, at temperature 0 where possible; Vectara's own Hallucination Evaluation Model (HHEM) judges each summary against its source. One task only: on Artificial Analysis's knowledge-and-abstention test (9 September 2026), GPT-6 Astra's hallucination rate fell from 92% for GPT-5.6 Sol to 51%, where that rate is $\text{incorrect} \div (\text{incorrect} + \text{partial} + \text{not attempted})$. The two measures count different things and are not comparable.

▼ Table view

Figure 3. Share of document summaries containing unsupported content, selected OpenAI models

Model (as listed by Vectara)	Hallucination rate in summaries (%)
gpt-5.4-nano (2026-03-17)	3.1%
gpt-6-sol	6.5%
gpt-6-astra	8.7%
gpt-5-nano (2025-08-07)	10.5%

Neither result is wrong. They measure different tasks, and a model's reliability on one does not settle its reliability on the other.

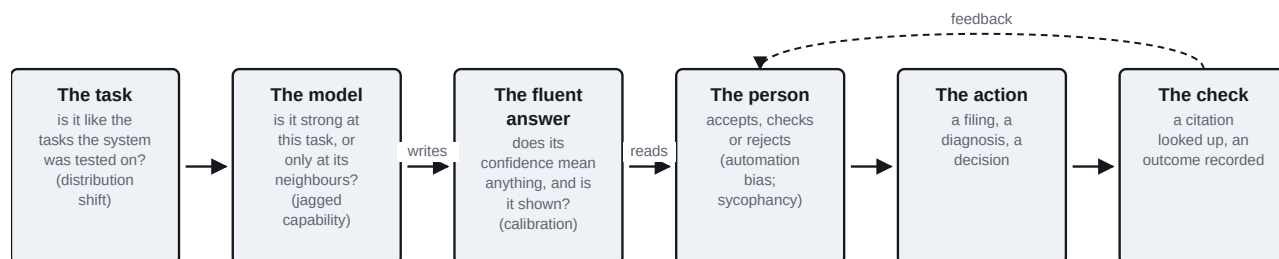
3.4 The person in the loop

A fluent answer does its damage only when someone acts on it, which makes the reader part of the system. In 1983 Lisanne Bainbridge, of University College London's psychology department, published "Ironies of Automation" in *Automatica*. Its central observation is that "the more advanced a control system is, so the more crucial may be the contribution of the human operator". Automation, she wrote, "by taking away the easy parts of his task", can "make the difficult parts of the human operator's task more difficult", and "perhaps the final irony is that it is the most successful automated systems, with rare need for manual intervention, which may need the greatest investment in human operator training". She also noted that people cannot sustain effective attention on a source of information "on which very little happens" for more than about half an hour.

The literature that followed named the resulting error **automation bias**. A 2010 review in *Human Factors* by Raja Parasuraman and Dietrich Manzey found that it "results in making both omission and commission errors when decision aids are imperfect", that it occurs "in both naive and expert participants", and that it "cannot be prevented by training or instructions". A related failure belongs to the machine side: a system that changes its answer to agree with a user who has offered no new evidence—sycophancy—hands the person's own error back with the authority of a second opinion.

The chain from question to consequence therefore has several links at which a fluent answer can go wrong, and each link has its own remedy.

Figure 4. Where a fluent answer can fail between the question and the consequence



Schematic, not measured data. Each node names the question that decides whether an answer can be relied on at that point; the return edge is the feedback without which a person's trust cannot become calibrated.

▼ Table view

Figure 4. Where a fluent answer can fail between the question and the consequence — stages

#	Stage	Note
1	The task	is it like the tasks the system was tested on? (distribution shift)
2	The model	is it strong at this task, or only at its neighbours? (jagged capability)
3	The fluent answer	does its confidence mean anything, and is it shown? (calibration)
4	The person	accepts, checks or rejects (automation bias; sycophancy)
5	The action	a filing, a diagnosis, a decision
6	The check	a citation looked up, an outcome recorded

Figure 4. Where a fluent answer can fail between the question and the consequence — connections

From	To	Label
The task	The model	
The model	The fluent answer	writes
The fluent answer	The person	reads
The person	The action	
The action	The check	
The check	The person	feedback

4. What happened, 2023–2026

The last three years supplied measurements at each link.

Date	Event	Link
15 March 2023	GPT-4 Technical Report: top-10% simulated bar exam; post-training "hurts calibration significantly"	calibration

Date	Event	Link
22 September 2023	Dell'Acqua et al.: +12.2% tasks inside the frontier, −19 points correct outside it	jagged capability
30 May 2024	Magesh et al.: legal research tools hallucinate 17%–33%, fewer than GPT-4	distribution shift; vendor claims
25 July 2024	Karpathy names "jagged intelligence"	jagged capability
10–12 July 2025	METR randomised trial: developers took 19% longer with AI, believing they were faster	the person's calibration
4–5 September 2025	OpenAI: accuracy-only scoring rewards guessing	calibration
3 February 2026	International AI Safety Report names an "evaluation gap"	distribution shift
9 February 2026	Bean et al., <i>Nature Medicine</i> : models alone right, people using them not	the person
9–22 September 2026	GPT-6 Astra: fewer hallucinations on one test, more than smaller siblings on another	jagged capability
25–28 September 2026	Massachusetts sanctions order; database reaches 2,095 decisions	all four

The METR trial deserves a fuller account, because it measured the person's calibration rather than the machine's. METR, a research group that evaluates AI systems, randomised 246 tasks from 16 experienced open-source developers, working in mature repositories they knew well, between allowing and forbidding AI. The tools were chiefly the Cursor editor with Anthropic's Claude 3.5 and 3.7 Sonnet models. (Disclosure: an Anthropic model was used in preparing this piece.) Before starting, the developers forecast that AI would cut their completion time by 24%; afterwards they estimated it had cut it by 20%; measured, "allowing AI actually increases completion time by 19%", with a confidence interval of 2% to 39% longer. METR listed what the result does not show, including that it does not claim its "developers or repositories represent a majority or plurality of software development work". On 24 February 2026 it reported a follow-up, begun in August 2025, whose raw estimates pointed the other way—18% less time for returning developers, 4% less for new ones, with intervals crossing zero—but judged that its new data gave "an unreliable signal", chiefly because a growing number of developers declined to take part rather than work without AI. The durable finding is the gap between what the developers felt and what the clock recorded, in early 2025, in that setting.

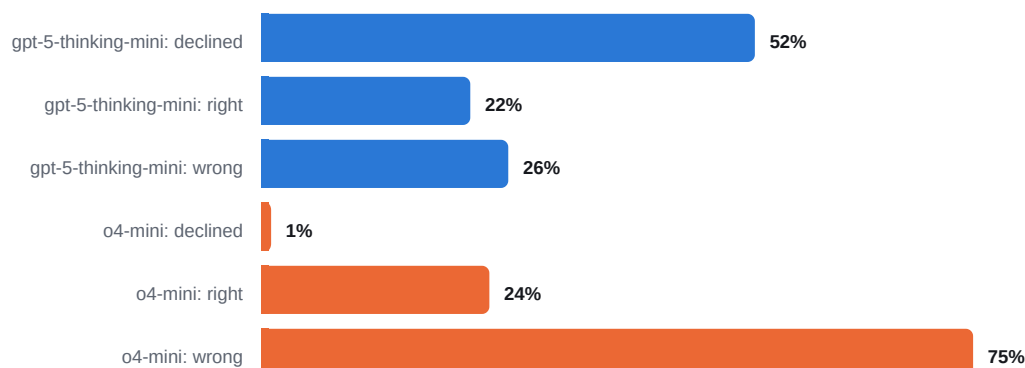
5. Two postures: fix the incentive, or test the pair

The first posture holds that the machine can be made to signal its own uncertainty if it is rewarded for doing so. In September 2025 four researchers, three of them at OpenAI—Adam Tauman Kalai, Ofir Nachum, Santosh Vempala and Edwin Zhang—argued that "language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty". OpenAI's accompanying blog post of 5 September was careful about causation: "Hallucinations persist partly because current evaluation methods set the wrong incentives. While evaluations themselves do not directly cause hallucinations," it continued,

most evaluations measure model performance in a way that encourages guessing rather than honesty about uncertainty.

That is Brier's complaint of 1950, restated for a new kind of forecaster. OpenAI's illustration, taken from its GPT-5 system card, compared two of its models on SimpleQA, a test of short factual questions.

Figure 5. Two OpenAI models on SimpleQA: answers declined, right and wrong (% of questions)



Source: OpenAI, 'Why language models hallucinate', 5 September 2025, citing the GPT-5 system card. Colour and label identify the model. Shares of all questions, not error rates among answered questions. OpenAI's own numbers about its own models; no independent audit.

▼ Table view

Figure 5. Two OpenAI models on SimpleQA: answers declined, right and wrong (% of questions)

Model and outcome	Share of questions (%)
gpt-5-thinking-mini: declined	52%
gpt-5-thinking-mini: right	22%
gpt-5-thinking-mini: wrong	26%
o4-mini: declined	1%
o4-mini: right	24%
o4-mini: wrong	75%

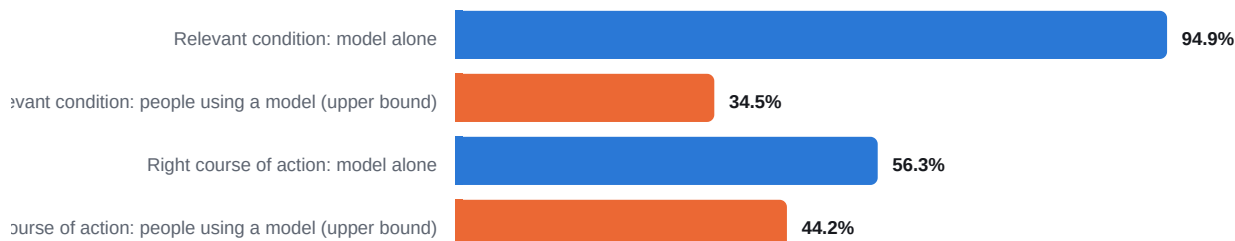
The older model was right slightly more often and wrong almost three times as often, because it almost never declined. On a leaderboard that counts only right answers, it would rank higher.

The second posture holds that the model's own behaviour, however well scored, is the wrong unit of measurement, because what is used in practice is a model and a person together. The clearest test is a randomised, pre-registered trial by Andrew Bean and colleagues at the University of Oxford, published in *Nature Medicine* on 9 February 2026. It gave 1,298 members of the British public one of ten medical scenarios written by doctors, and assigned them to consult GPT-4o, Llama 3, Command R+ or any source of their choice. The experiment ran from August to October 2024, so the models are not current ones. The abstract's result is stark:

Tested alone, LLMs complete the scenarios accurately, correctly identifying conditions in 94.9% of cases and disposition in 56.3% on average. However, participants using the same LLMs identified relevant conditions in fewer than 34.5% of cases and disposition in fewer than 44.2%, both no better than the control group.

On identifying conditions the control group did significantly better than those using the models; on choosing a course of action the differences were not significant.

Figure 6. The same models, alone and in the hands of the public (% of scenarios handled correctly)



Source: Bean et al., 'Reliability of LLMs as medical assistants for the general public: a randomized preregistered study', *Nature Medicine*, 9 February 2026, abstract. The abstract gives the participant figures as 'fewer than' 34.5% and 44.2% across the three models, so those bars are ceilings. Data collected 21 August to 14 October 2024 with GPT-4o, Llama 3 and Command R+; structured scenarios, not patients.

▼ Table view

Figure 6. The same models, alone and in the hands of the public (% of scenarios handled correctly)

Measure	Correct (%)
Relevant condition: model alone	94.9%
Relevant condition: people using a model (upper bound)	34.5%
Right course of action: model alone	56.3%
Right course of action: people using a model (upper bound)	44.2%

The authors drew the conclusion directly:

Standard benchmarks for medical knowledge and simulated patient interactions do not predict the failures we find with human participants.

The Massachusetts order adds a complementary duty rather than a test: the person using the tool remains responsible for the filing. It applies Federal Rule of Civil Procedure 11 to the lawyer whatever tool drafted his brief, and lists the range of sanctions other courts have imposed for fictitious or misleading citations—penalties, fee awards, referrals to disciplinary bodies, dismissal.

The two postures are complements rather than rivals. A system that declines or flags uncertainty when it should makes checking cheaper and targets it better. A person who knows where a system was tested knows where the checking is most needed. The criterion that fits the evidence is not how often a model is right in general, but whether the combination of model, task and reader produces fewer uncaught errors than the alternative.

6. What to watch

The Massachusetts case, 23 October and 9 November 2026. The order requires the parties to file a joint notice "on or before October 23, 2026" stating whether they have agreed the fees to be paid, and sets a status conference for 9 November.

The database's monthly count, end of October 2026. The baseline is 2,095 decisions as of 28 September. Monthly figures rather than the total are the useful comparison, since the latest weeks tend to fill in late. Whether the monthly pace falls below 100 as courts' warnings accumulate, or holds above it, is the observable question; either result could have more than one cause, since changes in use, scrutiny and reporting could each move the count.

The International AI Safety Report, if last year's timing holds. Its first "key update" of the last cycle appeared on 15 October 2025 and its second on 25 November 2025; its publications page lists no 2026 update yet. The February 2026 report named "an emerging 'evaluation gap': existing evaluation methods do not reliably reflect how systems perform in real-world settings". The test of any update is whether it reports measurements of systems in use, rather than only on benchmarks.

In February METR announced plans to redesign its developer-productivity study, without giving a date for results.

7. The idea to keep

Sounding right is not evidence of being right. A fluent answer shows that a system is good at producing fluent answers; its reliability depends on the task, on whether its confidence means anything, and on the person reading it. Three questions do most of the work. Is the task like the ones on which the system was tested, or has the world shifted? Does the system signal when it is unsure, and has anyone checked that the signal is calibrated? Does its competence at the neighbouring task say anything about this one? The standard reference remains Lazer and colleagues' three-page "The Parable of Google Flu" (*Science*, March 2014), which concerns search data rather than chatbots, and is the better guide for it.

Sources

Source	Date
U.S. District Court for the District of Massachusetts, Memorandum and Order on Motion for Sanctions, Civil Action No. 25-CV-12395-AK (Judge Angel Kelley)	25 September 2026
Damien Charlotin, <i>AI Hallucination Cases Database</i> (page and CSV), damiencharlotin.com/hallucinations	last updated 28 September 2026; read 30 September 2026
OpenAI, <i>GPT-4 Technical Report</i> , arXiv 2303.08774	March 2023
Magesh, Surani, Dahl, Suzgun, Manning and Ho, <i>Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools</i> , arXiv 2405.20362	30 May 2024
Ginsberg et al., <i>Detecting influenza epidemics using search engine query data</i> , <i>Nature</i> 457	19 February 2009

Source	Date
Lazer, Kennedy, King and Vespignani, <i>The Parable of Google Flu: Traps in Big Data Analysis</i> , Science 343	14 March 2014
Zech et al., <i>Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs</i> , PLOS Medicine	6 November 2018
Gwern Branwen, <i>The Neural Net Tank Urban Legend</i> , gwern.net/tank	2011, revised 2023
Glenn W. Brier, <i>Verification of Forecasts Expressed in Terms of Probability</i> , Monthly Weather Review 78(1)	January 1950
Andrej Karpathy, <i>Jagged Intelligence</i> , post on X	25 July 2024
Dell'Acqua et al., <i>Navigating the Jagged Technological Frontier</i> , Harvard Business School Working Paper 24-013	22 September 2023
Artificial Analysis, <i>Benchmarking GPT-6 Astra</i>	9 September 2026
Vectara, hallucination leaderboard	22 September 2026
Lisanne Bainbridge, <i>Ironies of Automation</i> , Automatica 19(6)	1983
Parasuraman and Manzey, <i>Complacency and bias in human use of automation</i> , Human Factors 52(3)	June 2010
Becker, Rush, Barnes and Rein (METR), <i>Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity</i> , arXiv 2507.09089	July 2025
METR, <i>We are Changing our Developer Productivity Experiment Design</i>	24 February 2026
Kalai, Nachum, Vempala and Zhang, <i>Why Language Models Hallucinate</i> , arXiv 2509.04664	4 September 2025
OpenAI, <i>Why language models hallucinate</i> (blog)	5 September 2025
Bean et al., <i>Reliability of LLMs as medical assistants for the general public: a randomized preregistered study</i> , Nature Medicine	9 February 2026
International AI Safety Report 2026, arXiv 2602.21012; publications page	3 February 2026; read 30 September 2026