

On 4 September 2026 Artificial Analysis, an independent firm that ranks AI models, dropped a graduate-level science test called GPQA Diamond from its headline index, describing it as an evaluation "that has now been saturated". Six days later Epoch AI, the research group behind the FrontierMath benchmark, announced that every problem in that test's hardest tier had now been solved by AI. A study published in the proceedings of this year's International Conference on Machine Learning (ICML) found, by its own measure, that 29 of 60 widely used tests of language models had largely lost the ability to separate the leading systems. A benchmark score is a claim that a small, fixed set of questions stands in for a skill, and that claim decays in three common ways—contamination, overfitting and saturation—each with its own history, its own signature and its own remedy.

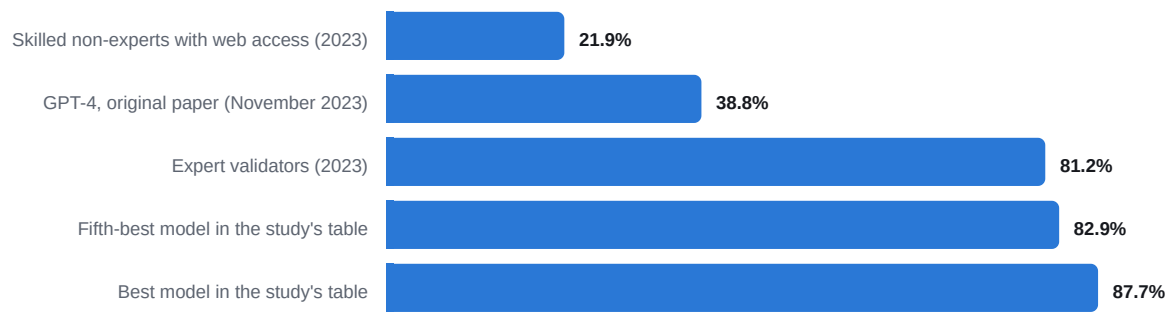
1. A test retired, a tier finished

Artificial Analysis publishes an "Intelligence Index" that folds a model's results on a basket of tests into a single number. Version 4.2, announced on 4 September, added two tests and removed one. The changelog gave the reason in a line:

GPQA Diamond, an exceptional scientific reasoning evaluation that has now been saturated

A post on the firm's account on X on 6 September gave a second reason beside the first: the test "provides little signal for frontier models given its saturation, and its multiple-choice format doesn't reflect real tasks in those domains". The firm still runs GPQA Diamond on new models and reports the result on its own; it has left the index, not the firm's test suite.

GPQA was published in November 2023 as "A Graduate-Level Google-Proof Q&A Benchmark" by eight researchers at New York University, two of whom also list affiliations with Cohere and Anthropic. Its Diamond subset holds 198 questions in biology, physics and chemistry, kept, with narrow exceptions, because both expert validators answered them correctly while most skilled non-experts, searching the web without restriction, did not. In the original paper GPT-4 answered 38.8% of the Diamond questions correctly. In the ICML study's table of leading scores, already present in its first version of February 2026, the five best models scored between 82.9% and 87.7%.

Figure 1. GPQA Diamond: from a hard test to a retired one (accuracy, %)

Sources: Rein et al., 'GPQA: A Graduate-Level Google-Proof Q&A Benchmark', arXiv 2311.12022, 20 November 2023, Table 5 (GPT-4 few-shot chain-of-thought; validator accuracies, which the authors mark as skewed by selection effects because the Diamond set was chosen using those validators' answers); Akhtar et al., 'When AI Benchmarks Plateau', arXiv 2602.16763 v4, 6 August 2026, Table 6 (top five models on the leaderboard the study used; the same rows appear, as Table 5, in the paper's first version of 18 February 2026, and the study prints no snapshot date). Artificial Analysis removed GPQA Diamond from its Intelligence Index on 4 September 2026. Questions have four answer options, so guessing scores about 25%.

▼ Table view

Figure 1. GPQA Diamond: from a hard test to a retired one (accuracy, %)

Who answered, and when	Share of the 198 Diamond questions answered correctly (%)
Skilled non-experts with web access (2023)	21.9%
GPT-4, original paper (November 2023)	38.8%
Expert validators (2023)	81.2%
Fifth-best model in the study's table	82.9%
Best model in the study's table	87.7%

Epoch's announcement, posted on 10 September, read: "Every FrontierMath Tier 4 problem has now been solved by AI, with GPT-6 Astra solving the last problem standing." Epoch counts a problem as solved once any model has solved it in any attempt (its October 2025 update reported "the total number ever solved"), which is not the same as one model scoring full marks. Two facts about the test belong beside the headline. Epoch's own page states that "FrontierMath was developed with funding from OpenAI, who has exclusive access to a subset of the benchmark"; in October 2025 Epoch put that subset at 28 of the 48 Tier 4 problems it then evaluated, holding back the other 20; the June revision later removed seven Tier 4 problems. The model that solved the last problem is OpenAI's. And on 12 June 2026, the page records, Epoch "released a major update, addressing errors in 42% of problems" across the benchmark's tiers.

The ICML paper, "When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation", is the work of 37 researchers led by Mubashara Akhtar of ETH Zurich and Anka Reuel of Stanford, assembled through a coalition called EvalEval. First posted in February and revised three times, most recently on 6 August, it examined 60 text-based benchmarks: those used in at least five of 61 model cards and technical reports published between January 2022 and November 2025 by developers including OpenAI, Anthropic, Google, Meta and Alibaba, supplemented by heavily cited benchmark papers and by hand-picked additions to cover particular designs. Its abstract reports that "nearly half of our benchmarks exhibit saturation, with rates increasing with age".

2. Two readings that mislead

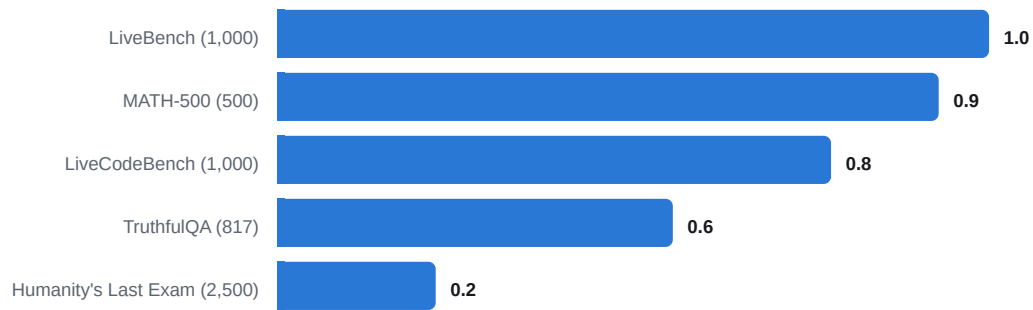
2.1 "Saturated means solved"

The natural reading of "saturated" is that the machines have learned the subject and the test has nothing left to ask. The study's definition is narrower:

A benchmark is saturated if the evaluated models cannot be reliably distinguished by their performance scores and any further improvements are not statistically distinguishable under the evaluation protocol.

In practice the authors take the five best scores on a benchmark's leaderboard, compare their spread with the uncertainty that a test of that size carries, and convert the result into an index between 0 and 1; an index of 0.7 or more counts as high saturation. The ceiling against which crowding is judged is "the highest observed model performance", not 100% and not a human score, and the authors note that "human-level performance does not imply saturation", because models can remain statistically distinguishable after passing a human baseline.

The consequence is that saturation can arrive well short of perfection. Of five benchmarks worked through in the paper's appendix, LiveBench, a test designed to resist contamination by changing its questions every month, scored 0.99, the most saturated of the five, with its leaders clustered within about a point of each other at around 79%. The authors read that as "model-level stagnation rather than task completion". Humanity's Last Exam, a deliberately harder test released in January 2025, scored 0.22.

Figure 2. How crowded the top of the leaderboard was: saturation index for five benchmarks

Source: Akhtar et al., 'When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation', arXiv 2602.16763 v4, 6 August 2026 (ICML 2026), Appendix E, Table 7. The index compares the spread of the five best leaderboard scores with the statistical uncertainty of a test that size; 0.7 and above is 'high', 0.9 and above 'very high'. Spread of the top five, in percentage points: LiveBench 1.09 (at about 79%), MATH-500 1.0 (98.2 to 99.2), LiveCodeBench 3.9, TruthfulQA 6.7, Humanity's Last Exam 11.4. The study prints no snapshot date; this table was added in a revision after its first version of 18 February 2026.

▼ Table view

Figure 2. How crowded the top of the leaderboard was: saturation index for five benchmarks

Benchmark (test questions)	Saturation index, 0 = top models well separated, 1 = indistinguishable
LiveBench (1,000)	1.0
MATH-500 (500)	0.9
LiveCodeBench (1,000)	0.8
TruthfulQA (817)	0.6
Humanity's Last Exam (2,500)	0.2

The authors attach a condition of their own. Saturation, they write, is "not always negative": "if the benchmark was valid", saturation "means that a task can be considered 'solved'". The conditional matters. A score supports a claim about a skill only to the extent that the test measures that skill, and a crowded leaderboard adds no evidence either way on that question. A saturated test of the wrong thing is a test of the wrong thing that everyone now passes. Saturation is also relative to a purpose: a test that no longer separates the strongest systems can still separate weaker ones, or catch a model that has regressed.

The headline count itself rests on a choice of yardstick. The paper's sensitivity analysis shows that the *ordering* of benchmarks from least to most saturated changes little when its two tuning parameters are varied (rank correlations of 0.88 to 0.92), but that only between 18% and 48% of benchmarks stay in exactly the same one of its five bands; the authors add that "most changes occur between neighbouring bins rather than large shifts". Together with a sample chosen for wide use rather than at random, that makes "twenty-nine of sixty" a reading of about half, on this study's yardstick, rather than a census of the field.

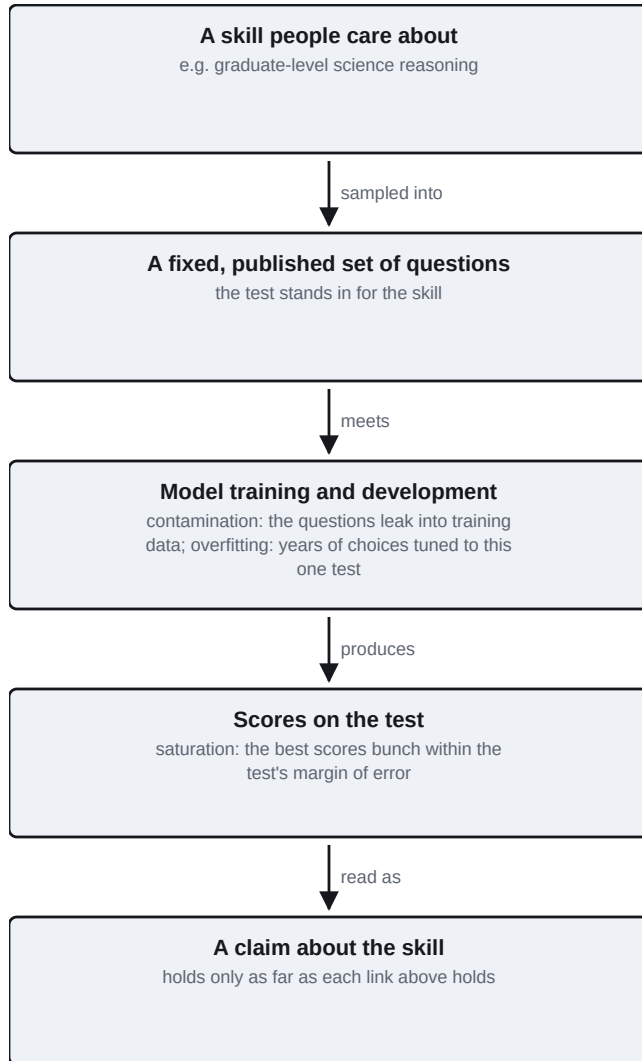
2.2 "So the scores are fake"

The opposite reading treats every high score as memorisation. The measured record does not support it as a rule. When OpenAI checked GPT-3 against its own training data in 2020, it found that potential contamination was "often high (with a quarter of benchmarks scoring over 50%)", yet "in most cases

performance changes only negligibly" when the overlapping questions were removed. When Scale AI wrote a fresh set of grade-school mathematics problems in 2024 to test for memorisation of an older set, it found real drops for some model families and "minimal signs of overfitting" for the models at the frontier. When researchers at Berkeley rebuilt the ImageNet test from scratch in 2019, the leading models lost accuracy but kept their order. A score that has rotted has not necessarily been faked; the useful question is which part of the claim still stands.

3. The idea: three ways a score decays

A benchmark score asserts that performance on a particular set of items stands in for a wider skill. The assertion has three weak points. The items may have leaked into what the model learned from. The field may have tuned its choices so closely to those particular items that the score no longer speaks for the skill. Or the items may have run out of power to tell the strongest systems apart. The first is contamination, the second overfitting, the third saturation.

Figure 3. Three points at which a benchmark score stops supporting the claim made for it

Schematic, not measured data. Each kind of decay breaks a different link, which is why each needs a different check: a fresh matched test for contamination, a re-collected test or a final private ranking for overfitting, and the spread of top scores against their uncertainty for saturation.

▼ Table view

Figure 3. Three points at which a benchmark score stops supporting the claim made for it — stages

#	Stage	Note
1	A skill people care about	e.g. graduate-level science reasoning
2	A fixed, published set of questions	the test stands in for the skill
3	Model training and development	contamination: the questions leak into training data; overfitting: years of choices tuned to this one test
4	Scores on the test	saturation: the best scores bunch within the test's margin of error
5	A claim about the skill	holds only as far as each link above holds

Figure 3. Three points at which a benchmark score stops supporting the claim made for it — connections

From	To	Label
A skill people care about	A fixed, published set of questions	sampled into
A fixed, published set of questions	Model training and development	meets
Model training and development	Scores on the test	produces
Scores on the test	A claim about the skill	read as

3.1 Contamination: the exam paper seen in advance

A student who has seen the exam paper can score well without knowing the subject. Language models learn from enormous scrapes of the internet, and benchmarks are published on the internet. The problem was recognised early. OpenAI's paper introducing GPT-3, posted on 28 May 2020, described an attempt to remove every benchmark's test questions from the training data, and then this:

Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model.

Unable to retrain, the authors measured the damage instead, rescored each benchmark on a "clean" subset of questions with no long overlap with the training data. Most scores barely moved; two results (PIQA and Winograd) were marked with asterisks, and several language-modelling benchmarks were dropped from the paper because almost all of their text appeared in the training set. The conclusion was carefully hedged: "either our conservative method substantially overestimated contamination or that contamination has little effect on performance".

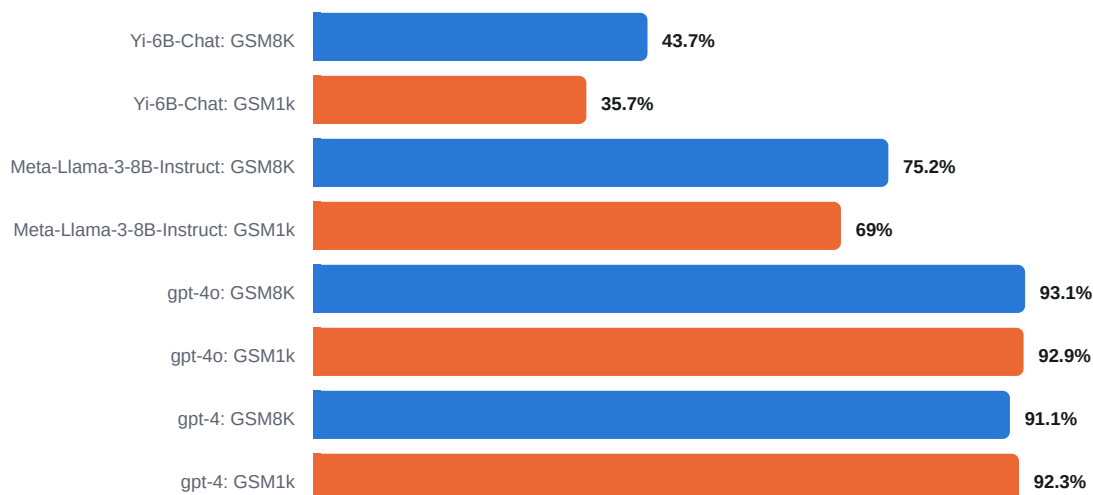
Good intentions on the benchmark side have not been enough either. BIG-bench, a large collaborative collection of test tasks, embeds a "canary" string whose purpose, its documentation says, is "to allow researchers to better filter BIG-bench tasks out of the training data for large language models". OpenAI's GPT-4 report of March 2023 states that "during our contamination check we discovered that portions of BIG-bench [48] were inadvertently mixed into the training set, and we excluded it from our reported results". A canary is a label and a detector, not a lock.

The cleanest way to see contamination is to date the questions. LiveCodeBench, introduced by researchers at Berkeley, MIT and Cornell in March 2024 and revised that June, collects programming-contest problems with their publication dates. Its authors observed "a stark drop" in the performance of a DeepSeek coding model on problems published after August 2023, just before the model's release, which they read as suggesting "that the earlier problems might indeed be contaminated". The drop appeared mainly on problems from one platform, LeetCode; on other platforms performance was "relatively smooth across the months".

Scale AI's GSM1k study, posted in May 2024 and revised in November, took the matched-fresh-test route. Its authors commissioned 1,205 new grade-school problems in the style of the widely used GSM8K test and found "accuracy drops of up to 8%"—eight percentage points in its appendix tables—with "several families of models showing evidence of systematic overfitting across almost all model sizes". (The first version of the paper said up to 13%; the figure was revised.) Models more likely to reproduce GSM8K's text verbatim

tended to show bigger gaps, a correlation the authors read as suggesting "that some models may have partially memorized GSM8k". They also found that "many models, especially those on the frontier, show minimal signs of overfitting".

Figure 4. The same kind of question, old and new: accuracy on GSM8K and on the freshly written GSM1k (%)



Source: Zhang et al. (Scale AI), 'A Careful Examination of Large Language Model Performance on Grade School Arithmetic', arXiv 2405.00332 v4, 22 November 2024, Appendix F, standard-prompt results (API models queried between 16 April and 10 July 2024; identifiers as the authors report them). Darker bars: GSM8K, the public test; lighter bars: GSM1k, 1,205 newly written problems matched in style and difficulty, which the authors describe as 'only highly similar, but not identically distributed'. The largest gap in the table is Yi-6B-Chat's 8.0 points; gpt-4 scored higher on the new set. A gap invites investigation; it does not by itself separate memorisation from other differences between the sets.

▼ Table view

Figure 4. The same kind of question, old and new: accuracy on GSM8K and on the freshly written GSM1k (%)

Model and test	Accuracy (%)
Yi-6B-Chat: GSM8K	43.7%
Yi-6B-Chat: GSM1k	35.7%
Meta-Llama-3-8B-Instruct: GSM8K	75.2%
Meta-Llama-3-8B-Instruct: GSM1k	69%
gpt-4o: GSM8K	93.1%
gpt-4o: GSM1k	92.9%
gpt-4: GSM8K	91.1%
gpt-4: GSM1k	92.3%

The International AI Safety Report, published in February 2026 under the chairmanship of Yoshua Bengio, put the general problem plainly: "many models may have been trained using data from these same benchmarks – a problem called 'data contamination', which most developers do not currently track or disclose".

3.2 Overfitting: fitting the sample, not the pattern

A model with enough freedom will find patterns in any sample, including patterns that are accidents of that sample. The defence is to judge the model on data it was not fitted to—a held-out test set. Benchmarks are held-out test sets shared by a whole field, which creates a subtler version of the problem. If many researchers try many ideas and keep the ones that score best on the same public test, the test is no longer truly held out: the field as a whole has been fitted to it. Researchers call this adaptive overfitting, "overfitting caused by test set reuse".

The fear is reasonable; the evidence on its size is more reassuring than the fear. In 2019 Benjamin Recht and three colleagues at the University of California, Berkeley rebuilt the test set of ImageNet, the image-recognition benchmark on which a decade of progress had been reported, by following the original collection procedure as closely as possible. Accuracy fell by 11 to 14 percentage points, a loss the authors equated to "approximately five years of progress". Yet the models kept almost exactly the same order, and

accuracy gains on the original test sets translate to larger gains on the new test sets.

Their results "suggest that the accuracy drops are not caused by adaptivity", but by the models' inability to handle slightly harder images than those in the original test. The authors add that "it is unclear what properties of our new images cause the accuracy drops". A companion study of 120 competitions on Kaggle, a platform where teams submit repeatedly against a public leaderboard before a final private ranking, found "somewhat surprisingly, little evidence of substantial overfitting". The fall on a fresh test was real; the evidence pointed away from years of tuning to the old test as its cause.

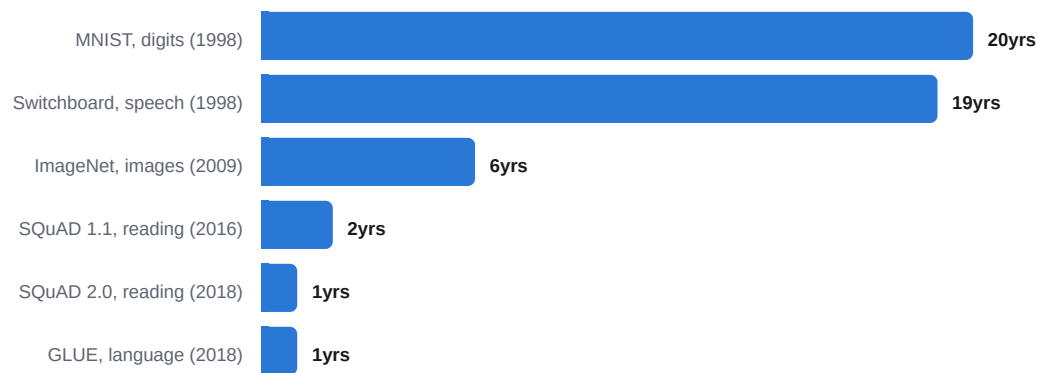
Whether that reassurance carries over to language models is open. The ImageNet and Kaggle studies examined test sets that were reused but did not leak into training data. Public language-model benchmarks face both problems at once, and GSM1k's family-specific gaps suggest that the combination can matter.

3.3 Saturation: a test that runs out of room

Every test has a ceiling. In 2021 Douwe Kiela of Facebook AI Research and colleagues described how quickly benchmarks now reach theirs:

While it used to take decades for machine learning models to surpass estimates of human performance on benchmark tasks, that milestone is now routinely reached within just a few years for newer datasets

Their figure traced six benchmarks from first result to the human estimate. Handwritten-digit recognition (MNIST) and conversational speech transcription (Switchboard) took about two decades; ImageNet took six years; the reading-comprehension test SQuAD took two; its successor, and the language-understanding suite GLUE, took about one.

Figure 5. Years from a benchmark's first plotted result to its first result at or above the human estimate

Source: Kiela et al., 'Dynabench: Rethinking Benchmarking in NLP', arXiv 2104.14337, 7 April 2021, Figure 1 ('Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero'). The paper prints the figure but not its data; the plotted points were recovered from the PDF's vector drawing and are accurate to about a year. MNIST and Switchboard begin at the plot's left edge, so their spans are lower bounds; MNIST sat just below the human mark from 2013 and crossed it at its next plotted point, 2018. 'Human performance' is each benchmark's own estimate, which Kiela et al. call 'the narrow criteria used to define human performance'.

▼ Table view

Figure 5. Years from a benchmark's first plotted result to its first result at or above the human estimate

Benchmark (first plotted year)	Years to reach the human estimate (approx.)
MNIST, digits (1998)	20yrs
Switchboard, speech (1998)	19yrs
ImageNet, images (2009)	6yrs
SQuAD 1.1, reading (2016)	2yrs
SQuAD 2.0, reading (2018)	1yrs
GLUE, language (2018)	1yrs

GLUE shows the cycle in its own documents. Its launch paper, posted in April 2018, concluded that "solving GLUE is beyond the capabilities of current models and methods". A year later the paper introducing its successor, SuperGLUE, reported that "performance on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research". On 6 January 2021 Microsoft announced that a single DeBERTa model had passed SuperGLUE's human baseline "for the first time in terms of macro-average score (89.9 versus 89.8)", and added that "the model is by no means reaching the human-level intelligence of NLU".

Two further features make saturation harder to read than a full score suggests. The first is that the ceiling is often set by the test's own errors. A re-annotation of MMLU, a widely used multiple-choice knowledge test, first posted in June 2024, estimates in its current version (January 2025) that "6.49% of MMLU questions contain errors", and found errors in 57% of the questions it analysed in the virology section; at the other end of the difficulty range, FrontierMath's June revision addressed errors in 42% of its problems. Near the top, a higher score can mean agreeing with a wrong answer key, and a correct answer can be marked wrong, so errors blur the ceiling rather than fixing it at a number. The second is that saturation drives replacement.

Humanity's Last Exam, released in January 2025, opens by observing that "LLMs now achieve over 90% accuracy on popular benchmarks like MMLU, limiting informed measurement of state-of-the-art LLM capabilities", and offers 2,500 expert-written questions "designed to be the final closed-ended academic benchmark of its kind".

4. What happened, 2018–2026

Date	Event	Decay it illustrates
April 2018	GLUE launched; "beyond the capabilities of current models and methods"	saturation
February 2019	ImageNet re-collected: accuracy down 11–14 points, order preserved	overfitting (largely not found)
May 2019	SuperGLUE launched because GLUE's human baseline had been passed	saturation
May 2020	GPT-3's filtering bug; most clean-subset scores barely move	contamination
January 2021	DeBERTa passes SuperGLUE's human baseline, 89.9 to 89.8	saturation
March 2023	GPT-4 report: parts of BIG-bench "inadvertently mixed into the training set"	contamination
November 2023	GPQA published; GPT-4 scores 38.8% on its Diamond set	—
March 2024 (revised June)	LiveCodeBench dates its problems; drops after models' cutoffs	contamination
May 2024 (revised November)	GSM1k: drops of up to 13% in its first version, up to 8% after revision; minimal at the frontier	contamination and overfitting
January 2025	Humanity's Last Exam, built because MMLU passed 90%	saturation
February–August 2026	"When AI Benchmarks Plateau": 29 of 60 benchmarks highly saturated	saturation
June 2026	FrontierMath v2 addresses errors in 42% of problems	the ceiling set by errors
4 September 2026	Artificial Analysis removes GPQA Diamond from its index	saturation
10 September 2026	Epoch: every FrontierMath Tier 4 problem solved by some AI	saturation

The ICML study's wider results fit the pattern, with a caution attached. Of its 60 benchmarks, 29 showed high or very high saturation, 14 of them very high. Among benchmarks released within the previous 24 months, 42.9% were saturated; among those more than 60 months old, 54.5%. The authors describe that trend as "modest and not statistically significant at conventional thresholds", though in their joint statistical model benchmark age and test-set size "show the most consistent effects". Larger test sets went with less saturation, and expert-written benchmarks were less saturated than crowdsourced ones of a similar age, though the authors note that curation categories also differ in age; two expert-curated tests, ARC-AGI and BIG-Bench Hard, "remain unsaturated despite prolonged exposure". How often a benchmark was cited did not predict saturation once its age was taken into account. The paper's own causal statement is hedged: its results are "consistent with" repeated optimisation compressing the differences between frontier models, "although our analysis does not directly identify the causal mechanism".

Figure 6. Sixty benchmarks, one yardstick: the ICML 2026 saturation study in four numbers

Source: Akhtar et al., 'When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation', arXiv 2602.16763 v4, 6 August 2026, published at ICML 2026: section 4 (counts and age bins), section 4.1 (public versus private), Table 1 (sensitivity to the parameters k and α). The sample is purposive, not random: benchmarks used in at least five of 61 model reports from January 2022 to November 2025, heavily cited benchmark papers, and additions chosen to cover particular designs. All four numbers are from version 4; the first version of 18 February 2026 counted 8 private benchmarks and had no sensitivity table. The study prints no snapshot date.

▼ Table view

Figure 6. Sixty benchmarks, one yardstick: the ICML 2026 saturation study in four numbers

Measure	Value
Benchmarks with high or very high saturation	29 of 60
Share saturated, released within 24 months	42.9%
Private test sets in the sample	4 of 60
Benchmarks keeping their band when the method's settings change	18–48%

5. Two postures: hide and refresh, or resolve and retire

The responses on offer divide into two, and they are aimed at different kinds of decay.

The first is to keep the questions out of reach of training, either by hiding them or by replacing them. Artificial Analysis's version 4.2 raised the share of its index weighted on private, held-out tests to 40%, "double the figure from v4.1", and said that "This reduces the ability for labs to game evaluations"; the firm says the share "will increase further in Index v5". Its current methodology table (version 4.3.2) marks tests carrying 45% of the weight as private, by a sum of the weights it lists; the page itself prints no total. LiveBench, introduced in June 2024, replaces about a sixth of its questions in each update, so that it is "fully refreshed roughly every 6 months", and withholds each month's new questions for a month. The current arXiv version of its paper (April 2025) calls it "Contamination-Limited", although the citation in the project's own README still reads "Contamination-Free"; its changelog entry of 25 November 2025 describes an update made "to resolve (to some extent) saturation and contamination of the benchmark". Microsoft's SWE-bench-Live, a test built from real software-repository issues, has promised since September 2025 that "each month, we will add 50 newly verified, high-quality issues to the dataset test split", while freezing two smaller splits so that leaderboard comparisons stay fair.

The second posture comes from the ICML study, and it is less comforting about the first. The four private benchmarks in its sample behaved like the 56 public ones:

Hiding test data does not appear to prevent saturation once benchmarks are widely adopted

Four is a small group, and the comparison is suggestive rather than conclusive. The worked examples in its appendix, chosen to span the scale rather than to rank designs, include a related caution: LiveCodeBench, the authors write, "demonstrates that also dynamically constructed benchmarks can saturate when evaluation resolution is limited". Their recommendations lead with resolution: larger test sets, harder items or scores broken down by sub-skill, so that "score differences between models exceed expected evaluation uncertainty"; confidence intervals and the spread of top scores reported beside the peak; and explicit criteria, written in at design time, for revising or retiring a test once the best systems can no longer be told apart. They endorse refreshing as well, but as one tool among several.

Private tests also carry a cost the public kind does not: outsiders cannot check them. FrontierMath is largely private, yet its funder holds a subset; Artificial Analysis's held-out tests cannot be inspected by the public. Moving questions out of view moves trust from the test to whoever holds it.

The criterion that fits the evidence is that the postures are complements, each aimed mainly at a different decay. Secrecy and freshness are aimed at contamination, and the study suggests that refreshing can also slow the convergence that comes with exposure. Size, difficulty and honest error bars are what let a test keep separating the strongest systems. Re-collected tests and final private rankings detect overfitting. A benchmark programme that does only one of these is defending against one failure while the other two proceed.

6. What to watch

Artificial Analysis's next index version, from October 2026. Its methodology page stood at version 4.3.2 on 1 October; the firm has promised further "incremental releases" and a version 5 in which the private share "will increase further". The checkable questions are whether the private share, 45% of the weight by the current table, keeps rising, whether another constituent is retired as saturated, and whether each retirement comes with its reason. Humanity's Last Exam, the least saturated of the study's worked examples, still carries 10% of the index's weight.

FrontierMath Erdős, Epoch's newer test, through October. Launched on 3 September with 68 unsolved "Erdős problems"—open questions associated with the mathematician Paul Erdős—whose solutions must be written as machine-checkable proofs in the Lean language, it began with GPT-6 Astra solving 2 of 68; Epoch says no earlier model solved any. The rate at which that number moves is the start of a new test's clock.

Whether the refreshing designs refresh, by early November. SWE-bench-Live's stated policy, since September 2025, is 50 new verified Python issues a month; its most recent dataset news, on 21 August 2026, reported 1,077 task instances in its multi-language edition. LiveBench's public changelog has no entry after 8 January 2026, although its code repository is still being updated. A "live" benchmark that stops adding questions becomes a static one, and decays like one.

7. The idea to keep

A benchmark score has a shelf life. Three questions establish how much of its claim is still standing. Could the model have seen these questions, or close copies of them, before it was tested? Has a whole field been tuning its choices against this one test for years? And can the test still tell the best systems apart by more than its own margin of error? None of the three makes a score worthless; each locates the part of it that has decayed. A thorough current treatment is Akhtar, Reuel and colleagues' "When AI Benchmarks Plateau" (ICML 2026), whose appendix walks through five benchmarks, from saturated to not, in a page; for the opposite caution, Recht and colleagues' "Do ImageNet Classifiers Generalize to ImageNet?" (2019) remains a clear demonstration that a falling score is not always a fraud uncovered.

Sources

Source	Date
Artificial Analysis, <i>Announcing Artificial Analysis Intelligence Index v4.2</i> , artificialanalysis.ai/articles	4 September 2026
Artificial Analysis, post on X on Intelligence Index v4.2	6 September 2026 (00:28 UTC)
Artificial Analysis, <i>Intelligence Benchmarking Methodology</i> , version 4.3.2	read 1 October 2026
Epoch AI, post on X on FrontierMath Tier 4	10 September 2026
Epoch AI, <i>FrontierMath Tier 4</i> benchmark page and <i>Benchmarks</i> hub	read 1 October 2026
Akhtar, Reuel, Soni, Ahuja et al., <i>When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation</i> , arXiv 2602.16763 v4 (ICML 2026)	18 February 2026; v4 6 August 2026
Rein et al., <i>GPQA: A Graduate-Level Google-Proof Q&A Benchmark</i> , arXiv 2311.12022	20 November 2023
Brown et al. (OpenAI), <i>Language Models are Few-Shot Learners</i> , arXiv 2005.14165	28 May 2020
OpenAI, <i>GPT-4 Technical Report</i> , arXiv 2303.08774	March 2023
BIG-bench, <i>training_on_test_set</i> README (canary GUID), in the BIG-bench repository on GitHub	read 1 October 2026
Jain et al., <i>LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code</i> , arXiv 2403.07974	12 March 2024; v2 6 June 2024
Zhang et al. (Scale AI), <i>A Careful Examination of Large Language Model Performance on Grade School Arithmetic (GSM1k)</i> , arXiv 2405.00332	1 May 2024; revised November 2024
International AI Safety Report 2026, arXiv 2602.21012	February 2026; arXiv 24 February 2026
Recht, Roelofs, Schmidt and Shankar, <i>Do ImageNet Classifiers Generalize to ImageNet?</i> , arXiv 1902.10811 (ICML 2019)	13 February 2019
Roelofs et al., <i>A Meta-Analysis of Overfitting in Machine Learning</i> , NeurIPS 2019	December 2019
Kiela et al., <i>Dynabench: Rethinking Benchmarking in NLP</i> , arXiv 2104.14337	7 April 2021
Wang et al., arXiv 1804.07461, <i>GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding</i>	20 April 2018

Source	Date
Wang et al., <i>SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems</i> , arXiv 1905.00537	2 May 2019
Microsoft Research, <i>Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark</i>	6 January 2021
Gema et al., <i>Are We Done with MMLU?</i> , arXiv 2406.04127 (v3)	6 June 2024; v3 10 January 2025
Phan et al., <i>Humanity's Last Exam</i> , arXiv 2501.14249 (quoted from v11)	24 January 2025; v11 28 July 2026
White et al., <i>LiveBench: A Challenging, Contamination-Limited LLM Benchmark</i> , arXiv 2406.19314; LiveBench changelog	27 June 2024; changelog read 1 October 2026
Microsoft, <i>SWE-bench-Live</i> README, in the SWE-bench-Live repository on GitHub	read 1 October 2026