When the test is *already in the training data*

How researchers learned to measure how much of a benchmark score comes from a model having seen the test, and where the evidence still disagrees.

MITWA LAB, MUMBAI

IN PLAIN WORDS

A benchmark is a fixed set of test questions used to compare AI systems. If a system met those questions, or close copies, while it was being built, its score tells you less about what it can do. That problem is called contamination. This review follows how researchers learned to measure it, from a 2019 experiment that rebuilt two image tests, through methods that can spot a memorised test without seeing the training data, to newly written tests that mirror old ones. The effects are real and uneven. On one fresh maths test some model families lose several points while frontier models barely move, and the main studies read a drop in different ways. Preparing a model for a known kind of test, which some researchers regard as legitimate, can also distort comparisons.

A test set rebuilt from scratch

Overfitting to a test set does not need a leaked file. When many researchers choose models by their scores on the same test set, the winners may be the ones that happen to suit that set. Recht et al. (2019) checked this worry on two image benchmarks, CIFAR-10 and ImageNet, by repeating each dataset's original collection process as closely as they could and assembling new test sets of 2,000 images (CIFAR-10) and 10,000 images (ImageNet). Across a broad range of models, accuracy fell by 3 to 15% on CIFAR-10 and by 11 to 14% on ImageNet.

If heavy tuning were the cause, the most heavily tuned models should have fallen furthest, and the pattern did not fit that expectation. The ranking of models was almost exactly preserved, and every point gained on the old test set was worth more than a point on the new one. The authors concluded that adaptivity was an unlikely cause and that the new images were slightly harder. Keeping only the easiest candidate images recovered the original ImageNet accuracies almost exactly. Three sampling strategies applied to one pool of candidate images, with average annotator selection frequencies of 0.73, 0.85 and 0.93, produced ImageNet test sets whose top-1 accuracy differed by about 14% between the two extremes.

A drop on fresh data is therefore not by itself evidence of memorisation, because the new test may simply be harder, and on ImageNet the way a test was assembled moved scores by about as much as the drop.

Seen is not the same as used

The language-model version of the question was harder to ask, because the training text of many of the largest models is not public. Magar and Schwartz (2022) made it answerable at small scale by training BERT models of two sizes themselves, on Wikipedia text with chosen test examples and their labels added, and comparing accuracy on test examples the model had seen against matched ones it had not. They separate memorisation, meaning the model can fill in a seen example's label, from exploitation, meaning that memory shows up as better accuracy on the task afterwards. With a downsampled sentiment set inserted 200 times, memorisation reached 60% and exploitation almost 40%, yet in another setting 15% memorisation came with less than 1% exploitation. The corpora held up to 600 million

tokens, and the authors warn that the results may not carry over to larger models.

Sainz et al. (2023) then set out the vocabulary in a position paper. They distinguish exposure of the annotation guidelines, of the raw text a benchmark was built from, and of the labels. Exposure of the evaluation labels invalidates an experiment. Exposure of the training or development splits leaves comparisons with models built on those same splits intact, but it invalidates conclusions that claim zero-shot or few-shot performance. Exposure can happen during pre-training, during instruction tuning, and after deployment, when a service improves from what its users send. They said outright that the extent of the problem was unknown, since it is hard to measure.

Detection when the training data is secret

Oren et al. (2023) built a test that needs only the probability a model assigns to a piece of text. Many benchmarks are exchangeable, meaning the order of their examples carries no information, so a model that never saw one should find its published order and a random shuffle equally likely. A model that saw it in the published order tends to prefer that order, and the test turns the preference into a p-value with a controlled false-positive rate. The guarantee would break if a real benchmark contained order-dependent elements such as sequential indexes or duplicates, so the authors inspected each dataset by hand and also ran the test on BioMedLM, a model trained only on PubMed data, as a negative control. To check the method, they trained a 1.4 billion parameter model from scratch on about 20 billion tokens of Wikipedia text, into which ten test sets of roughly 500 to 1,500 examples each had been inserted at known repetition counts. Sets inserted 10 or more times produced vanishingly small p-values, and three sets inserted once were not flagged (p-values of 0.16 to 0.48 on the paper's main test). In a separate canary study with repetition counts of 1, 2, 4 and 7, detection began at around four insertions. A clean result from this test therefore rules out heavy repetition and says little about a single copy.

Their audit of four public models on eight benchmarks found little evidence of pervasive contamination. The one flagged result, ARC for Mistral-7B ($p = 0.001$), sits at the boundary of significance once multiple tests are allowed for, and the authors ask for caution. A heuristic analysis of MMLU gave small p-values for LLaMA2-7B and Mistral-7B, which the authors describe as consistent with earlier reports of mild MMLU contamination.

Golchin and Surdeanu (2024) take the opposite route and ask the model to continue a benchmark item. The request names the dataset and split and gives the start of an item, and a partition is flagged when, within a sample of ten items, at least one completion matches the real remainder exactly or at least two match it nearly. In their best variant, GPT-4, shown a few labelled examples, judges the matches. Across 14 dataset partitions drawn from seven datasets, this variant agreed with two human experts in all 14 cases for GPT-4 and in 13 for GPT-3.5. That is agreement with human judgement of the same completions and says nothing certain about the training data. The authors also contaminated GPT-3.5 on purpose, by continued training on items from three datasets, to check the thresholds they use, and each produced at least one exact match.

A third family uses release dates. Li and Flanigan (2024) compared 12 models on 16 classification datasets, splitting each model's datasets into those released before its training data was collected and those released after. With difficulty controlled, the models did better on the earlier datasets, and on classification tasks released later they rarely beat a majority-class baseline, by a statistically significant margin. The authors say the approach has a high chance of catching contamination that is there but low precision, because other things change over time as well, whereas direct inspection and extraction are precise and miss a great deal. Roberts et al. (2023) applied the same logic to problems from Codeforces and Project Euler, using the September 2021 training cutoff shared by GPT-3.5-Turbo and GPT-4. Before the cutoff, a problem's pass rate rose with how often it appeared on GitHub. After the cutoff that association was not significant at the 0.05 level (on Codeforces, $p = 0.988$ for GPT-4 and $p = 0.081$ for GPT-3.5-Turbo), and none of the models solved any of the 73 Project Euler problems released later.

Exposure through ordinary research use

Balloccu et al. (2024) pointed to a quieter route. As they read OpenAI's data usage policy, content typed into the ChatGPT web interface could be used for training, while content sent through its API or business products was not, though they note that before 1 March 2023 the policy allowed API data to be used as well. They identified 255 papers that evaluated GPT-3.5 or GPT-4 in the first year after release, judged 212 of them relevant, and counted the benchmark samples that researchers had sent in a way that could be used for training. Their estimate of the total was about 4.7 million samples from 263

benchmarks, and about 42% of the relevant papers (90 of 212) had sent such data.

The authors had not yet tested what the exposure did to later scores, and they say their list is incomplete because it rests on what papers made public.

Writing the test again, by hand

Zhang et al. (2024) commissioned GSM1k, 1,205 grade-school maths problems written wholly by human annotators without help from any language model, to mirror GSM8k, a widely used test of the same kind. Nineteen annotators who had not written problems tried to spot the one GSM1k problem among five and found it 21.83% of the time across 1,205 attempts, against 20% by chance. Two models that predate GSM8k, GPT-NeoX-20B and GPT-2, scored about the same on both.

On leading models, accuracy fell by up to 8%, with the Phi and Mistral families showing a drop at almost every release and size, while frontier models showed minimal signs of overfitting. Across models, the likelihood of generating GSM8k test items rose with the size of the gap (Spearman rank correlation 0.36, $p = 0.03$), which the authors read as partial memorisation. They also say contamination is probably not the whole explanation, since some outliers do not fit, and that heavily overfit models such as Phi-2, with a 6% drop, still solved over half of the new problems. They withheld GSM1k to keep it clean and promised to release it by June 2025 at the latest.

Zhu et al. (2024) went further and generate evaluation items on demand from graph-structured procedures with adjustable difficulty, instead of collecting a fixed set. Some models that claimed large gains on existing benchmarks did poorly on the generated items, which the authors read as a sign of possible low-quality training data or contamination.

When a drop has a second explanation

Mirzadeh et al. (2025) rewrote 100 GSM8K questions as symbolic templates and generated 50 sets of variants in which names and numbers change while the reasoning stays the same. Across 25 models, the original GSM8K score sat on the high side of the spread of variant scores for 21 of them, which the authors say could reflect contamination. The drops were larger for several small open models, such as Gemma2-9B and the Phi-3 family, and negligible for Llama3-8b and GPT-4o. When they added one clause that looks relevant but changes nothing, all the models tested fell, by over 65% for Phi-3-mini. For that result the authors suggest that the models' reasoning is mostly based on pattern matching rather than formal reasoning. For the gap between the original and symbolic scores they offer contamination as one explanation.

Zhang et al. (2024) read their own evidence differently. They lean toward the view that sufficiently strong models learn elementary reasoning, so that a model can have seen GSM8k and still generalise. The two papers do not strictly conflict, since one compares a test with a hand-written copy and the other measures how scores react to small edits, and both leave partial memorisation as one contributor. Read together, they suggest a limit on inference: a gap on a fresh test could contain memorisation, extra difficulty and sensitivity to surface wording, and neither experiment separates these within a single model.

Preparing for a kind of test

Dominguez-Olmedo et al. (2025) argue that part of the distortion comes from practices they do not count as malpractice. They use the term training on the test task for any use of knowledge about the evaluation tasks at training time, such as adding instruction-style data to pre-training text or choosing training data by how well it scores on benchmarks, and they treat it as acceptable or at least unavoidable. Across 56 base models of 70 million to 70 billion parameters, models released after November 2023 outscored older ones at equal training compute by over 7 points on MMLU and 19 points on GSM8K, differences that were statistically significant. When every model was then trained further on the same task-relevant data, roughly 100,000 multiple-choice examples taken from other benchmarks' training sets or roughly 600,000 maths examples, the gap became small and not statistically significant. In a controlled version, further training of the older models alone on the test task recreated differences of the kind seen between newer and older ones. The authors take this as suggesting that newer models mainly differ from older ones in how much they trained on the test task.

Recasting ARC and HellaSwag in the multiple-choice format of MMLU reproduced the same pattern, which the authors take to suggest that the gains reflect skill at that format more than recall of test items. They state that the adjustment needs substantial extra training and that suitable task data may be costly or unavailable. The causal reading also rests on assumptions set out in an appendix, and the main comparison is between existing models grouped by training date rather than models assigned to groups.

What is not yet known

Detection at the level of a single copy is open. The shuffling test of Oren et al. (2023) could not flag a benchmark inserted once, and the authors released their trained models so that others could attempt it. Until a method works at that level, a clean result from that test rules out heavy repetition and little else. Li and Flanigan (2024) say something similar of direct inspection and extraction: finding nothing by those routes does not show that a benchmark was unseen.

Balloccu et al. (2024) did not test whether exposure through ordinary research use changed later scores. Magar and Schwartz (2022) measured how far memorisation turns into better task accuracy only in BERT models trained on corpora of up to 600 million tokens. Zhang et al. (2024) and Roberts et al. (2023) saw patterns linked to memorisation in large models, but could not inspect the training data, so neither could compare seen with unseen examples.

None of the studies cited here divides a drop on a fresh test between memorisation, difficulty and sensitivity to wording within one model family. The rank correlation of 0.36 in Zhang et al. (2024) is relatively low, and the authors point to outliers that contamination alone does not explain. Mirzadeh et al. (2025) built their variants from only 100 questions.

Fresh tests also age, because a withheld set such as GSM1k stays fresh only while it stays private. Zhang et al. (2024) note that evaluating proprietary models meant sending the data through an API, which they believe providers typically do not train on, and they held out a small number of items in case of leakage. The sources read here do not measure how fast a withheld set loses value, and they do not say whether GSM1k was released on the promised schedule.

The line between legitimate preparation and contamination is also drawn by definition. Dominguez-Olmedo et al. (2025) treat training on the test task as acceptable, while Sainz et al. (2023) treat exposure of the training and development splits as a threat to zero-shot claims.

References

Balloccu, S., Schmidtova, P., Lango, M., and Dusek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. arXiv:2402.03927. https://arxiv.org/abs/2402.03927

Dominguez-Olmedo, R., Dorner, F. E., and Hardt, M. (2025). Training on the test task confounds evaluation and emergence. arXiv:2407.07890 (v3). https://arxiv.org/abs/2407.07890

Golchin, S., and Surdeanu, M. (2024). Time travel in LLMs: Tracing data contamination in large language models. ICLR 2024. arXiv:2308.08493. https://arxiv.org/abs/2308.08493

Li, C., and Flanigan, J. (2024). Task contamination: Language models may not be few-shot anymore. AAAI 2024 (arXiv version read). arXiv:2312.16337. https://arxiv.org/abs/2312.16337

Magar, I., and Schwartz, R. (2022). Data contamination: From memorization to exploitation. arXiv:2203.08242. https://arxiv.org/abs/2203.08242

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2025). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. ICLR 2025. arXiv:2410.05229. https://arxiv.org/abs/2410.05229

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T. B. (2023). Proving test set contamination in black box language models. arXiv:2310.17623. https://arxiv.org/abs/2310.17623

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? arXiv:1902.10811. https://arxiv.org/abs/1902.10811

Roberts, M., Thakur, H., Herlihy, C., White, C., and Dooley, S. (2023). Data contamination through the lens of time. arXiv:2310.10628. https://arxiv.org/abs/2310.10628

Sainz, O., Campos, J. A., Garcia-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., and Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. arXiv:2310.18018. https://arxiv.org/abs/2310.18018

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., et al. (2024). A careful examination of large language model performance on grade school arithmetic. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2405.00332. https://arxiv.org/abs/2405.00332

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. (2024). DyVal: Dynamic evaluation of large language models for reasoning tasks. ICLR 2024. arXiv:2309.17167. https://arxiv.org/abs/2309.17167