What a test of a given size *can prove*

One benchmark of 500 questions, followed through the arithmetic: how precise a single score is, and how small a gap between two systems it can detect.

MITWA LAB, MUMBAI

IN PLAIN WORDS

A benchmark is a set of test questions, and a system's score is the share it gets right. The questions are a sample of those that could have been asked, so part of any score is luck, and another 500 questions of the same kind would give a slightly different number. This essay follows one test of 500 questions through the sums. For a score near 70%, a result on such a test is uncertain by about four points either way. Comparing two systems question by question on the same test is sharper than comparing their totals, by an amount depending on how often the systems agree. Questions in related groups, and the luck of training, blunt the comparison again. We argue that a reported comparison should say how small a difference its test could have detected.

Five hundred questions and one score

LongMemEval, a benchmark for the long-term memory of chat assistants, has 500 questions covering five abilities, one of which is declining to answer when the history holds no answer (Wu et al., 2025). We borrow its size, and only its size, for the running example. Every system and score below is invented for illustration and describes no real model.

Miller (2024) supplies the frame the rest of this essay relies on. The questions in an evaluation are treated as draws from a much larger, unseen population of questions that exercise the same skill, so a score is an estimate with a standard error like any other estimate from a sample. Suppose, as an illustration, that system A gets 350 of the 500 questions right. Its score is 70%, and a 95% interval for its accuracy on the wider population runs from about 66% to 74%, which is eight points of width from a test most people would call large.

That interval says how far the score would move if the 500 questions were drawn again from the same population, and nothing about questions gathered some other way. Recht et al. (2019) rebuilt the test sets of CIFAR-10 and ImageNet by following the original collection procedures closely. Accuracy fell by 3 to 15% on the first and 11 to 14% on the second, far outside any sampling interval, although the models kept almost exactly the same order; the authors trace the drop to new images that were slightly harder.

The uncertainty grows on small parts of a test. LongMemEval has 30 abstention questions (Wu et al., 2025), and if system A, again for illustration, declined correctly on 27 of them, the textbook interval of the score plus or minus 1.96 standard errors would run from 79% to just over 100%, which includes impossible values. Bowyer et al. (2025) show by simulation that intervals built on the central limit theorem contain the truth less often than they claim on small tests (at 100 items, nominal 95% intervals covered the true value 92.5% of the time), while the Wilson score interval and simple Bayesian intervals held their stated coverage. For 27 of 30 the Wilson interval runs from about 74% to 97%. Miller (2024) builds on the central limit theorem, and Bowyer et al. (2025) dispute its use below a few hundred items. For 350 of 500 the two kinds of interval differ by less than half a point, but Bowyer and colleagues find that intervals built on the theorem also fall short of their stated coverage when two models are compared and when questions are grouped, so the dispute bears on the subsets, on scores near 0 or 100%, and on the grouped example later in this essay, where 500 questions carry the information of about 179.

A four-point lead, analysed two ways

Now add system B, which in the illustration gets 370 of the same 500 questions right, a score of 74%. If the two scores are treated as independent estimates, the standard error of the difference combines both standard errors, and the 95% interval for B's lead runs from 1.6 points in A's favour to 9.6 points in B's, wide enough that the test cannot tell the two systems apart.

The two systems sat the same test, though, and that changes the arithmetic. Systems tend to find the same questions easy and the same ones hard, so their results are correlated, and an analysis of the question-by-question differences removes the shared part of the noise (Miller, 2024). For right-or-wrong scoring the paired analysis is McNemar's test, which looks only at the questions on which the two systems disagree (Card et al., 2020). Suppose, still for illustration, that B is right and A wrong on 50 questions, A right and B wrong on 30, and the systems agree on the remaining 420. These are the same two scores as before and the lead is still 4 points, but the paired interval now runs from about 0.5 to 7.5 points in B's favour, and the exact McNemar test gives $p \approx 0.03$.

How much pairing gains depends on how often the systems agree. Using outputs from many real systems on several language tasks, Berg-Kirkpatrick et al. (2012) found that systems with more similar outputs reach significance with smaller gains in the metric, because their errors are more correlated. The illustrative pair agrees on 84% of questions; with less agreement the paired interval would widen toward the unpaired one.

Questions that come in groups

Everything so far assumes that each question is an independent draw. Many tests ask several questions about one passage or one conversation, and questions that share a source have correlated answers, so the test holds fewer independent observations than it has questions and a standard error that ignores the grouping comes out too small. Miller (2024) recommends clustered standard errors, which account for the groups, and reports them for real evaluations run on Anthropic's models, where the clustered standard errors were 1.10 to 3.05 times the naive ones, with the largest ratio on DROP, a reading test that asks several questions about each passage.

LongMemEval builds a separate history for each question (Wu et al., 2025), so it does not ask several questions about one passage as DROP does, although its 30 abstention questions were made by modifying questions of the other types. Suppose, for illustration, that the 500 questions had come as 50 conversations of ten questions each, with a modest correlation of 0.2 between the paired differences inside a conversation. The standard correction for grouped samples then multiplies the variance by 2.8, the 500 questions carry about as much information as 179 independent ones, and the paired interval for B's lead widens to run from 1.8 points in A's favour to 9.8 in B's, so it once more includes zero.

When the system is itself a draw

The analysis above treats each system as fixed and only the questions as random. That fits a comparison of two finished systems, and fails when the claim is about a method or a training recipe, since running the same recipe again would produce a slightly different system. Madaan et al. (2024) trained ten language models of the same design that differed only in the random seed used to initialise them. Averaged over the course of training, the standard deviation across seeds was 1.11 points on HumanEval (164 problems), 2.15 on COPA (100 items) and 0.80 on ARC-C (1,165 items), generally smaller than the 95% sampling margin on the same test but well above zero.

Reimers and Gurevych (2017) report a starker case, a named-entity recognition system that was rerun 86 times with only the seed changed and scored between 89.99 and 91.00 F1 on the same test; the gap between its best and worst runs was itself statistically significant. Seeds are not the largest source of such noise, either. Bouthillier et al. (2021) varied each source of randomness separately in five case studies and found that resampling the data the model was trained and tested on mattered most, with weight initialisation usually contributing less than half as much variance.

To see the effect in the example, suppose for illustration that each system's score moves by one point, as a standard deviation, from one training run to the next. The difference between a single run of each then carries extra noise of about 1.4 points, and the paired interval widens to run from 0.5 points in A's favour to 8.5 in B's. A fixed system that samples its answers carries a related noise of its own, from the sampling. For that case Miller (2024) suggests scoring several sampled answers per question, and argues against lowering the temperature to quiet the noise, since that changes the system being measured.

How many questions a lead needs

Statistical power is the probability that a test detects a difference that is really there (Card et al., 2020). In the paired illustration, with the systems agreeing on 84% of questions, a true 4-point lead would be found significant at the 5% level about 61% of the time with 500 questions under the normal approximation used for the intervals above (about 57% with the exact McNemar test), and an unpaired analysis would find it about 29% of the time. Under the same approximation the smallest lead the paired test would catch 80% of the time is about 5 points, and with agreement held at 84%, catching a lead 80% of the time takes about 780 questions for 4 points, about 3,100 for 2 points and about 12,500 for 1 point. On these illustrative assumptions, a test of a few hundred questions is too small to check a claimed gain of one or two points.

Miller (2024) reaches the same order of magnitude with inputs he describes as fictional but reasonable, about 969 questions to detect a 3-point difference 80% of the time, and suggests that new evaluations hold at least 1,000 questions. In real parsing data, Berg-Kirkpatrick et al. (2012) found that a gain that would probably reach $p < 0.05$ was about 3.00 F1 on a 100-sentence slice of the Brown corpus and 0.22 F1 on the whole corpus of roughly 24,000 sentences.

Card et al. (2020) set published practice against this arithmetic. On MRPC, a paraphrase task with 1,725 test items, they estimated that beating the best model of the time with 80% power needed a gain of 1.62 points, while papers claiming the top result reported gains over their baselines of 0.63 points on average. For machine translation, they found that a test of 2,000 sentences has about 75% power to detect a difference of one BLEU point. A test with little power also distorts the results it does report as significant, because those results are more likely than usual to overstate the effect or to get its direction wrong.

The case for testing on fewer items

The strongest opposing view comes from work on making evaluation cheaper. Polo et al. (2024) used item response theory, a family of models from educational testing, to pick 100 representative questions from MMLU's roughly 14,000 and to turn the answers into an estimate of the full score. On models held out from the selection, the estimate from those 100 items was within about 2 points of the full-benchmark score on average, and in one analysis of 99 models no error exceeded 4 points apart from a single very weak model. If 100 well-chosen items can stand in for 14,000, it is fair to ask why a comparison needs thousands.

The two claims concern different quantities, and an average error of 2 points in estimating a score is too coarse for a comparison of two systems whose claimed gap is itself one or two points. The method also depends on the patterns of right and wrong answers seen in earlier models, and Polo et al. (2024) expect larger errors for a model whose pattern differs, for example one that fails easy questions while solving hard ones. Madaan et al. (2024) tried item analysis and item response theory as ways to reduce benchmark variance and found that they struggled to reduce it meaningfully. When they scored their 70 models on the 100-item subsets of Polo et al. (2024), the estimated averages held up fairly well, but the standard errors were much larger and up to 12% of pairs of models came out in the opposite order from the full benchmark. The authors advise against such subsets for evaluations meant to compare models.

What a reported comparison should carry

In two surveys, most papers reported no significance test, and in the more recent one most also gave no uncertainty estimate. Dror et al. (2018) surveyed the experimental long papers of ACL 2017 and found that 63 of 180 included a significance test, and that 21 of those 63 did not name the test used. Bean et al. (2025) reviewed 445 benchmarks for language models and found statistical tests or uncertainty estimates in 16.0% of them.

Our argument is that a reported comparison should state, beside the difference, the smallest difference its test could have detected. That one number takes in the number of items, whether the analysis was paired, how the items were grouped, and whether the claim is about a finished system or a method. A reader who sees a 4-point lead next to a minimum detectable difference of 5 points knows what the lead can carry. The number has limits of its own. It rests on inputs that are themselves estimates, such as how often the two systems agree, and it says nothing about whether a result survives questions gathered another way, which Recht et al. (2019) showed is a separate matter.

Our technical notes 01 and 03 apply the zero-failure case of this arithmetic to the lab's own tests of 400, 143 and 60 items. Note 03 treats one kind of comparison of two builds, a clean result in which the smaller build is never worse; this essay extends the same arithmetic to comparisons in which both systems get some questions wrong.

What is not yet known

- How much pairing gains when a model judge does the scoring. LongMemEval's judge agreed with expert judgments on 97 to 98% of answers on average, in a check of 30 questions per type (Wu et al., 2025), but how a judge's errors correlate across the two systems being compared is not reported.
- How often published systems agree item by item. That rate sets the power of a paired test; Card et al. (2020) estimated it by regression on the test predictions of ten high-performing models.
- How far beyond the small sizes that Bowyer et al. (2025) simulated the gap between interval methods persists, particularly when items are both grouped and few.
- Whether the seed variance measured for one family of models trained from scratch (Madaan et al., 2024) carries over to larger models, or to changes made after the main training run.
- Whether a selected subset in the manner of Polo et al. (2024) can be built to preserve the power to compare two systems; the existing subsets did not in the tests of Madaan et al. (2024), which reversed up to 12% of pairwise comparisons.

References

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., et al. (2025). Measuring what matters: Construct validity in large language model benchmarks. NeurIPS 2025 Datasets and Benchmarks Track. arxiv.org/abs/2511.04703

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In Proceedings of EMNLP-CoNLL 2012, 995–1005. aclanthology.org/D12-1091/

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., et al. (2021). Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3 (MLSys 2021). arxiv.org/abs/2103.03098

Bowyer, S., Aitchison, L., and Ivanova, D. R. (2025). Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In Proceedings of ICML 2025, PMLR 267. arxiv.org/abs/2503.01747

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. (2020). With little power comes great responsibility. In Proceedings of EMNLP 2020. arxiv.org/abs/2010.06595

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In Proceedings of ACL 2018, 1383–1392. aclanthology.org/P18-1128/

Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., and Hupkes, D. (2024). Quantifying variance in evaluation benchmarks. arXiv:2406.10229. arxiv.org/abs/2406.10229

Miller, E. (2024). Adding error bars to evals: A statistical approach to language model evaluations. arXiv:2411.00640. arxiv.org/abs/2411.00640

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. (2024). tinyBenchmarks: Evaluating LLMs with fewer examples. In Proceedings of ICML 2024, PMLR 235. arxiv.org/abs/2402.14992

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In Proceedings of ICML 2019. arxiv.org/abs/1902.10811

Reimers, N., and Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Proceedings of EMNLP 2017, 338–348. aclanthology.org/D17-1035/

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. (2025). LongMemEval: Benchmarking chat assistants on long-term interactive memory. ICLR 2025. arxiv.org/abs/2410.10813