# MOST HIGHLIGHTED COMPARISONS CANNOT BE CHECKED FROM THE PRINTED PAGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Results tables in machine learning often signal the strongest method by boldfacing a value, but such claims can be independently assessed only when the paper reports sufficient uncertainty information. We study how often this is possible using random samples totaling 2895 papers accepted at NeurIPS and ICML in 2024 and 2025. An automated pipeline identifies highlighted results and their strongest competitors, interprets table structure and metric direction, and determines whether each comparison contains enough information for statistical testing. In the 2025 sample, only 16.9% of highlighted comparisons are testable from the reported uncertainty and run counts. Among these testable comparisons, 47.6% of highlighted differences are not statistically distinguishable from zero using Welch's $t$-test at the stated run counts, and this fraction increases after correcting for multiple comparisons within tables. The 2024 sample exhibits the same overall pattern, and sensitivity analyses across alternative statistical and reporting assumptions do not change the qualitative conclusion. Validation against manually checked reference sets supports the reliability of the extraction pipeline, and the resulting data and evaluation artifacts are released with the paper. Our findings suggest that uncertainty and run counts should be reported at the level of the comparisons used to support empirical claims.

## 1 INTRODUCTION

A results table makes a claim. One number is set in bold, the others are not, and the reader is invited to conclude that the bold method is better. The claim is quantitative, so it can be wrong in a quantitative way: the difference may be smaller than the noise, or the largest of many differences examined at once.

Whether a reader can tell is a separate question from whether the claim is true, and it is the question we ask. Re-testing a difference between two reported means needs three quantities: a dispersion on each value and the number of runs the two rest on. Miss any one and the comparison cannot be re-tested from the printed page, by a reviewer, by a reader deciding what to build on, or by the authors a year later. Other routes to a check, a $p$-value or a marker on the difference, exist and are counted too (Section 8); they are rare.

We measure how often those three things are present. Across 1798 papers accepted at NeurIPS and ICML in 2025, 78.6% of highlighted comparisons carry no uncertainty on either side and 16.9% can be tested; 1097 papers from 2024 give the same picture. We ask whether the printed evidence is enough for a reader to check a comparison, which is a question about reporting rather than about whether the results are true, and one that is answered by looking at the page.

**Contributions.** Prior audits of reporting practice annotate papers by hand, which bounds them to a few hundred papers and a label per paper (Table 7, Appendix F). We work at the unit a results table claims in, the comparison.

- **A claim unit recoverable from typography.** The object of measurement is the bolded value against the best unbolded value in its group, and that value, the metric's direction and the table's orientation are all recovered from where the bold falls, so one pipeline covers whole venues whatever a benchmark measures.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Checkability on 32300 recovered comparisons**, about half the census an image reader finds on the same pages. 78.6% carry no error bar on either side as read, 65.6–77.5% after the instrument's measured recall; 16.9% can be tested. A paper-level audit cannot see this: one error bar in an appendix counts as reporting variability while the main claim stays unchecked.
- **Three questions a paper-level label cannot answer.** Whether the claimed difference survives the authors' own error bars, among the comparisons that print what a test needs (47.6% do not, 60.6% after correcting for the rest of its table); whether it is larger than the noise other papers report at that run count (48.7% are not, a design of Christodoulou et al. (2025) carried out by machine); and whether it is significant on the benchmark's own split (46.9% are not), a question Card et al. (2020) asked of a handful of benchmarks by hand.
- **A reference table for what benchmarks can resolve** (Appendix B), from published split sizes alone: AIME supports no claim below 33 points, a 500-item benchmark none below 8. The table is meant to be read while an experiment is being designed.
- **A checklist "yes" is not a printed error bar.** Of 657 NeurIPS papers answering that they report error bars, or other appropriate information about statistical significance, 411 print no error bar in any results table; 30.4% of those justify the answer by having used several seeds, which is averaging, not reporting.
- **An instrument scored against reference sets**, field by field, on rendered tables read with its values hidden, boxed and unboxed, on random pages read end to end, and against arXiv LaTeX source; released with the per-comparison database and the renderings.

## 2 What a results table claims

Let a *comparison group* be a set of cells the authors intend to be read against each other: a column, when methods occupy rows, and a row, when methods occupy columns. Within a group, exactly one cell is typically set in bold. That cell is the claim; the best cell that is not bold is the counterfactual the reader is invited to reject.

Two properties needed for analysis are recoverable from the typography alone, which makes the method independent of what the metric means. *Direction*: if the bold cell is the largest value in its group, higher is better; if the smallest, lower; if neither, the group is not a simple best-of comparison and is discarded rather than guessed. *Orientation*: some tables put methods in rows and benchmarks in columns, others the reverse, and reading one as the other compares a method against itself on a different benchmark. Whichever grouping yields more groups containing exactly one bold cell is the one the authors were marking.

A comparison is *testable* when both values carry a dispersion and the number of runs is stated, or the dispersion is itself a standard error or interval. We compute Welch's two-sided $t$ from the printed quantities alone, each standard error being the printed deviation over $\sqrt{n}$ and the degrees of freedom Satterthwaite's, in the normal limit where no run count is printed, and apply Holm–Bonferroni within each table, the family being the testable comparisons in that table (median 5), since the table is the unit at which the authors chose what to bold; Table 2 gives the paper, and every group in the table, as family. A bold cell tied with an unbolded one claims nothing and is dropped (785 groups). Welch's $t$ is what a reader with the same numbers should compute; a two-sample $z$, which takes a deviation from three runs as exact, is reported alongside (Section 5). Both cells are maxima of noisy values, and the winner is the maximum of more of them, so under any one convention the printed gap is if anything inflated and the failure rate computed on it if anything low. Two assumptions run the other way: independence, which overstates the rate wherever seeds were matched (Section 8), and a copied rival, whose run count the page does not state (Section 9).

## 3 Corpus

We sample from the open proceedings of NeurIPS and ICML: 1798 papers from 2025 (1059 NeurIPS, 739 ICML) and 1097 from 2024, each a simple random sample without replacement from that year's two accepted-paper lists, with a fixed seed; of the 2000 drawn for 2025, 126 could not be fetched and 76 exceeded 30 MB, so the population is accepted papers whose PDF could be fetched

Table 1: The same measurement in two samples, a year apart, not pooled. Percentages of comparisons are of all highlighted comparisons in that sample. Brackets give 95% intervals from a bootstrap that resamples *papers* rather than comparisons, because one author's habits govern every table in a paper: the intraclass correlation is 0.27, so 5465 comparisons carry about as much information as 811 independent ones.

| | **2024** | **2025** |
|---|---|---|
| Papers | 1097 | 1798 |
| Highlighted comparisons | 14741 | 32300 |
| *Can the comparison be checked?* | | |
| Neither value has an error bar | 79.3% [74.8, 83.2] | 78.6% [75.1, 82.0] |
| after the instrument's measured recall | | 65.6–77.5% |
| Error bar given, runs not stated | 4.0% | 2.6% |
| Testable | 15.2% [11.6, 19.1] | 16.9% [13.9, 20.2] |
| *Of those that can be checked* | | |
| Not distinguishable from zero, Welch's $t$ | 48.5% [40.6, 56.7] | 47.6% [42.4, 52.5] |
| after the instrument's measured verdict error | | 42.8–52.9% |
| Not distinguishable, Holm-corrected | 59.4% [51.4, 67.4] | 60.6% [54.5, 66.4] |
| after the instrument's measured verdict error | | 54.5–64.6% |
| Not distinguishable, two-sample $z$ | 41.6% | 40.2% |
| Median gap, failing | 2.06% | 1.74% |
| Median gap, surviving | 12.69% | 9.13% |
| *At the level of whole papers* | | |
| Print an explicit error bar | 25.5% | 26.0% |
| State a number of runs | 43.7% | 42.8% |

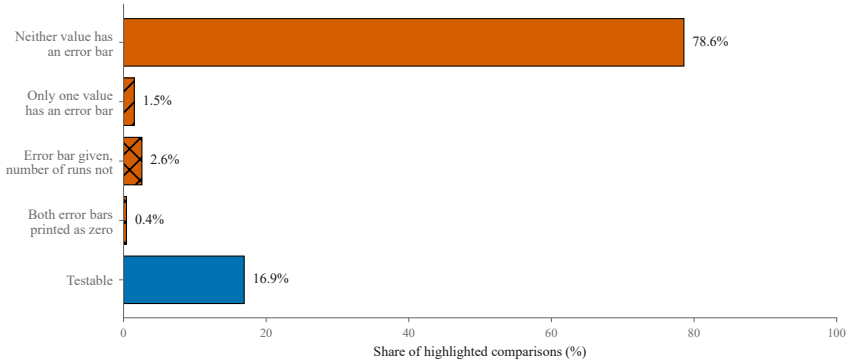


Figure 1: What stops a highlighted comparison from being checkable. The overwhelming majority print no uncertainty on either side; the fully testable slice is the small one at the bottom.

and read (Appendix K). The years are two samples, not pooled; the second was processed after the extractor was final, so it tests whether the result holds across a year, not the instrument; ICLR was not drawn.

## 4 How much can be checked

Table 9 (Appendix K) gives the funnel from PDF to analysed comparison. The population we measure is the tables that mark a winner and, within those, the groups where the mark is a claim of victory rather than a label on the authors' own row: right for the question, but filtered. Whatever the filter selects for, it is not the absence of uncertainty: tables that survive it carry an error bar 23.0% of the time, the tables it discards 10.4% (Section 9 reads a sample of them).

Figure 1 gives the decomposition. Of 32300 highlighted comparisons in 2025, 78.6% have no error bar on either value, 1.5% have one on only one side, 2.6% carry error bars whose number of runs

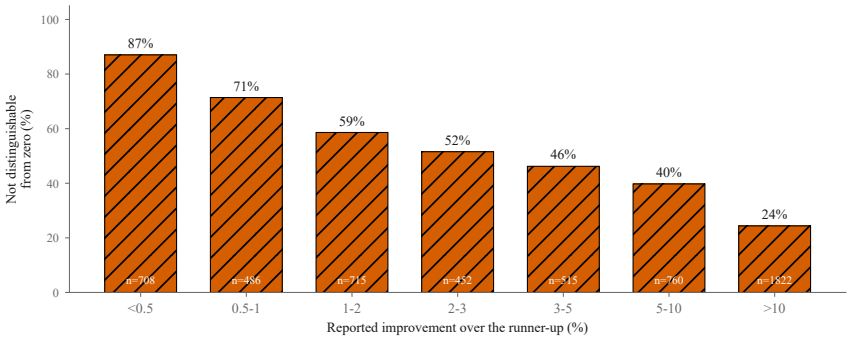


Figure 2: Within each band of reported improvement, the share that cannot be distinguished from zero using the authors' own error bars. Failures concentrate in the small margins, where leaderboard positions are decided.

is never stated, and 0.4% carry bars that are both zero or interquartile ranges, leaving 16.9%. The 2024 sample gives 79.3% and 15.2%, within a point and inside either bootstrap interval.

The pattern is not one community's, though its size is: by subfield the share without an error bar runs from 58.0% in reinforcement learning to 89.6% in language-model papers (Appendix G).

Of whole papers, 26.0% print an explicit error bar in any results table and 42.8% state a number of runs anywhere. Among testable comparisons the stated run count has median 5 (quartiles 3 and 6) and 27.5% rest on three runs or fewer.

## 5 What survives

Of the 5465 testable comparisons in 2025, 47.6% [42.4, 52.5] are not distinguishable from zero at the $5\%$ level using the numbers the authors printed. Holm–Bonferroni within each table raises this to 60.6% [54.5, 66.4]; the correction moves 13.0% [11.0, 15.0] of testable comparisons from significant to not, a paired quantity estimated far more precisely than either level.

A failure here says that the printed evidence does not support the claim at the stated run count; whether the claim itself is true is a separate question. At three runs a test has little power, and that is the evidence a reader of such a table is given. A two-sample $z$ on the same comparisons, each printed deviation taken as exact, fails 40.2%, 48.1% after Holm within table: the reading most favourable to the field among readings of what the page prints; only the matched-seed hypothetical of Section 8 goes lower. The gap between the two is the run count: Welch's critical value at three runs is $2.78$ against $1.96$, and 27.5% of testable comparisons rest on three runs or fewer. Table 2 carries both readings.

All intervals resample papers rather than comparisons (2000 draws, percentile intervals). With an intraclass correlation of 0.27 for the fail-or-survive indicator and 22.3 testable comparisons per paper, the design effect is 6.74: treating comparisons as independent would report intervals 2.6 times too narrow.

Figure 2 shows where the failures sit, and the pattern is monotone: 87% of comparisons reporting an improvement below half a point cannot be distinguished from zero, against 24% of those reporting more than ten; failing improvements have a median relative size of 1.74% against 9.13% for those that survive: the margins that decide leaderboards are the ones that cannot be supported.

## 6 Is the gap larger than typical noise at that run count?

The next question is whether a reported gap is large relative to the noise other papers report at that run count. Christodoulou et al. (2025) asked it of MICCAI papers by hand, imputing a dispersion where none was printed and estimating the correlation between methods rather than assuming it; this is the first half of that design carried out by machine. The noise level must not come from the paper

being judged: a threshold computed from its own error bar is its test statistic rescaled, and reading it as a separate finding is the observed-power fallacy (Hoenig & Heisey, 2001). We therefore take the run count from the paper and the noise from everyone else. Each value is bucketed by magnitude, $(0, 1]$ or $(1, 100]$; the noise level is the median standard deviation printed by other papers on the same benchmark in that bucket where a caption names one and ten cells from three papers exist (9.3% of comparisons), else on the paper's subfield (90.7%), else across the corpus. The statistic is the gap over the imputed standard error of the difference, reported below Welch's critical value at the stated run count, the headline's footing. It is a comparison against typical noise, not a test on the paper's own numbers, and the noise is crude: one median per subfield and scale, whatever the metric.

It can be asked of every comparison with a stated run count and two values in one bucket, error bars or not: 13509 comparisons from 462 papers. At a median of 5 runs, 48.7% of reported gaps fall below that cut on the imputed noise; on the 8360 comparisons that print no error bar at all, 50.1%, the first figure in this paper that covers that population, and the least secure: the noise is transferred from papers that print deviations to papers that do not, and the two need not share it; and a run count read from elsewhere in the paper may belong to another experiment: on 40 bar-less comparisons read with the whole paper the attached count is the table's in 21 of 40, every count from the caption or page (4 of 4) but only 17 of 36 from elsewhere. Restricted to caption-or-page counts the shares are 47.3% on 2918 comparisons and 44.3% on 1198 bar-less ones. On the 4645 testable comparisons the imputed noise puts 45.8% below the cut and the paper's own test 49.5%, with per-comparison agreement of 70.1% ($\kappa = 0.40$; a constant verdict would agree in 50.5%): the rates are close, the individual verdicts often differ, as a crude noise level should produce (Appendix G by subfield).

## 7 How large a test set would these gaps need?

Everything so far takes the uncertainty that matters to be variance across training runs, which is why a comparison needs an error bar and a run count to be testable. For an accuracy on a fixed test set the dominant uncertainty is over examples instead, and a binomial interval follows from the accuracy and the number of test items alone; the missing quantity is then the test-set size. Given the gaps the papers report, how large would a shared test set have to be for those gaps to be distinguishable from zero? Restricting to comparisons whose two values both lie in $[40, 100]$ gives 12359 comparisons from 621 papers, 2.3 times the base of the variance-based analysis, because it needs no error bars. F1, BLEU and mean IoU live in that range too and are not proportions over items, so we also report the stricter subset whose caption or column header names an accuracy-like metric: 4426 comparisons from 296 papers.

The median reported gap is 1.55 points, and detecting it at 80% power would require about 10,097 test items; 78.0% of these comparisons would need a test set larger than GSM8K, and 45.3% one larger than MMLU. On the accuracy-like subset the median gap is 1.51 points and the requirement 9,328 items (77.5% above GSM8K): the filter changes the population, not the answer. Treating the two values as independent binomials is conservative, since systems scored on the same items are positively correlated, so the true requirement is at most what we state. Widening the range filter to $(0, 100]$ moves the GSM8K share by a point (Appendix I).

This is a different question from Section 5: seed variance is the noise behind a claim about a training procedure, item variance the noise behind a claim that a system is better at the task; most papers make the second claim in words and the first in their error bars, when they print any.

**Against the benchmark actually used.** Where a caption names exactly one benchmark, the size of that benchmark's evaluation split is public, and the smallest difference it can resolve follows from that size and the two accuracies being compared, computed here at their pooled value. On 2217 such comparisons from 162 papers, the median reported gap is 1.2 points against a median resolution of 1.65. Two thresholds follow from the split: $1.96$ standard errors of the difference, below which a gap is not significant at the $5\%$ level, and the minimum detectable difference at 80% power, below which the benchmark could not be relied on to detect it. 46.9% of the gaps fall below the first and 54.7% below the second. Treating the two systems as independent makes both thresholds coarser than they are, so these are upper bounds; the systems really are scored on the same items, and under paired scoring the shares are 39.6% and 47.0% at $\rho = 0.5$, and 24.3% and 30.7% at $\rho = 0.9$.

Table 2: Both headline figures under each decision the pipeline had to make, each row on a stated population. The span across conventions is far wider than the sampling interval in both cases, and wider still for checkability than for the failure rate.

| | $n$ | Uncorrected | Holm |
|---|---|---|---|
| As reported: Welch's $t$, Holm within table | 5465 | 47.6% | 60.6% |
| Two-sample $z$, deviations taken as exact | 5465 | 40.2% | 48.1% |
| One-sided | 5465 | 40.3% | 56.4% |
| Unlabelled bars read as standard errors | 5465 | 55.3% | 67.2% |
| Run count from the caption or the table's page | 2370 | 41.7% | 53.5% |
| Holm family the paper, not the table | 5465 | 47.6% | 66.7% |
| Holm family every highlighted comparison in the table | 5465 | 47.6% | 61.2% |
| Only bars whose meaning is stated | 3536 | 47.8% | 61.3% |
| Run count taken from the caption only | 1760 | 39.0% | 50.4% |
| Tie-marked comparisons removed | 5452 | 47.5% | |
| Reweighted to the population's subfield mix | 5465 | 45.4% | |
| Reweighted to the population's gap distribution | 5465 | 46.5% | |
| Every copied rival removed, both extremes | | 44.7%–50.2% | |
| Sampling interval, conventions held fixed | 5465 | 47.6% [42.4, 52.5] | 60.6% [54.5, 66.4] |
| *Matched-seed hypothetical, not a reading of the page* | | | |
| Compared values correlated at $\rho = 0.5$ | 5465 | 40.5% | 54.3% |
| Compared values correlated at $\rho = 0.9$ | 5465 | 30.3% | 42.0% |
| *And the share that is checkable at all* | | | |
| As reported | 32300 | 16.9% | |
| Run count taken from the caption only | 32300 | 5.4% | |
| Unlabelled bars read as standard errors | 32300 | 19.5% | |
| Any marker or $p$-value counted as a check | 32300 | 26.8% | |
| Run counts scaled by their measured precision | 32300 | 15.1% | |
| Run counts the pattern misses, restored at the audited rate | 32300 | 18.1% | |
| Accuracy on a benchmark of known size (Section 7) | 32300 | 6.9% | |
| Either route | 32300 | 23.3% | |

Table 4 (Appendix B) gives the breakdown for every benchmark with thirty or more matched comparisons. The share below resolution is not monotone in split size, because it also depends on the gaps a field reports: WinoGrande and TruthfulQA, with around a thousand items each, have almost nothing else, since a thousand items cannot resolve the one-point differences reported on them, while AIME's thirty items resolve a third of their gaps because the gaps reported on AIME are tens of points. This is a second route to a check, open to any accuracy on a named benchmark whether or not the paper prints a bar: 6.9% of all comparisons can be checked this way, and 23.3% by either route (Table 2).

## 8 How much is a definitional choice?

Several decisions in the pipeline are not forced by the data, and they matter more than sampling does. Table 2 gives the failure rate under each, with the population each row is computed on. Over the rows that share the full population of 5465 testable comparisons the span is 40.2% to 55.3%, against a bootstrap interval of 47.6% [42.4, 52.5] that captures sampling alone: definitional uncertainty is larger than the statistical kind.

The headline reads unlabelled bars as standard deviations, which favours the field, takes the two values as independent, which does not, and uses Welch's $t$ at the stated run count; the subset where authors state what their bar means fails at 47.8%. The span is wide because the papers do not print which convention they used.

**Where a run count comes from, and what a bar means.** A caption sometimes states how many runs produced its means; more often the count appears once in the setup, and the pipeline then takes the smallest count stated anywhere in the paper (Appendix K). Counting only captions gives 5.4% testable and 39.0% failing; a count stated anywhere gives 16.9% and 47.6%, which we report.

By source the rate is 40.7% on the 931 comparisons whose count is in the caption, 48.5% on the 712 from the table's page, 51.1% on the 3424 from elsewhere in the paper and 31.9% on the 398 standard errors and intervals that need none: the paper-wide count, 62.7% of the testable set, is the one convention whose direction we cannot sign. Of the papers that print an explicit error bar, only 70.0% state a run count anywhere. Nor do most say what the bar is: reading every unlabelled bar as a standard error rather than a standard deviation moves the failure rate to 55.3% (Appendix I); we use the deviation reading, the more conservative.

**Whether the two values are independent.** Our test treats the winner and the runner-up as independent samples of run-to-run variation; matched seeds or shared initialisations would correlate them, and print cannot say whether they were. Table 2 therefore carries the paired reading, with $\mathrm{Var}(A - B) = \sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B$ at $\rho = 0.5$ (40.5%) and $\rho = 0.9$ (30.3%), as a matched-seed hypothetical that reads nothing off the page; a paired re-test proper needs per-seed differences no table prints.

**Checkability moves further than survival does.** The share that can be checked is the more convention-dependent headline. A standard error is already the dispersion of a mean, so a comparison whose bars are standard errors needs no run count: reading every unlabelled bar that way makes 19.5% of comparisons testable rather than 16.9%, while demanding the run count come from the caption leaves 5.4%, a factor of 3.6 between the ends. Other routes to a check are counted per table: markers on values, $p$-values or the word "significant" near the table, and interval cells $a\,[l, u]$. Of the comparisons with no error bar on either side, 12.6% sit in a table with a marker of some kind and 5.9% with a $p$-value or the word significant (a dagger often means "copied", so the first is a ceiling); crediting all of them raises the checkable share from 16.9% to 26.8%, and 20.9% of the papers with no error bar in any table mention error bars, shading or seeds in a figure caption, which the instrument cannot see. The claim in our title survives every reading: at the most generous end, three highlighted comparisons in four cannot be checked from what the table prints.

## 9 Threats to validity

**The extractor could be wrong, so it was scored.** This is the threat that matters: whether the pipeline read the right two cells, the right two error bars and the right run count. So we built a reference set (Table 5, Appendix D): draws of claims and tables, each rendered from the PDF with the pipeline's chosen cell boxed and nothing else, and read with the pipeline's values hidden. The reader recorded the winner, the rival, both means and bars, the bar kind, and the run count found in the paper's seed, run and trial sentences, supplied in full. The reader was the vision-language model named in the AI use statement: these figures are agreement between two extractors of different technology, not precision against a human reference, and the box anchors the reader on the pipeline's choice of winner, though not on its values. No usable human labels exist (Appendix D records the one attempt), and the renderings, labels and scoring script are released (`https://anonymous.4open.science/r/checkable--E2C5`) so that anyone can make them.

The development set of 100 testable comparisons, read against the pipeline as it stood after the first six bugs of Appendix H, found five more; the pre-fix pipeline agreed with the reader on the whole claim (the bold cell, its rival and all four values) in 87 of 97 readable items (89.7% [82.1, 94.3]). A held-out draw of 50 from papers the first set never touched agreed on the claim in 40 of 50 (80.0% [67.0, 88.8]) and on the verdict in 31 of 32, and found two more. After the fixes the same items agree in 89 of 94 and 36 of 41, figures optimistic in the way a set that chose the fixes must be; Appendix D prints every scoring.

A fourth draw was made after the released pipeline was tagged, from papers no earlier set touched, and nothing was fixed on it: 100 testable comparisons from 50 papers, with 40 no-bar and 40 no-run-count items. On it the pipeline and the reader agree on the whole claim in 69 of 99 (69.7% [60.0, 77.9]; resampling papers, [51.2, 87.0]), 11 of the 30 disagreements coming from one paper whose bold marks every valid method rather than a winner; on the verdict in 66 of 68 items both could test, the reader failing 54.4% and the pipeline 51.5% of them; and the pipeline fails 44.0% of all 100, against 47.6% in the population. Of the no-bar items 39 of 40 carry none, the paper states a run count somewhere for 19 of 40 of the no-run-count items (Table 2 restores them), and the runner-up is a copied number in 5 of 97. Read again with nothing marked, so that the box cannot

anchor it, the reader lists the pipeline's winner among the single-bold groups it finds in 85 of 100, with the same rival in 78 of 85 of those (78.0% [68.9, 85.0] of all); two further readers, a smaller model of the same family and one from another vendor, agree with the first on the rival in 54 of 58 and 54 of 57 of a sixty-item subset and with the pipeline on the verdict in 28 of 29 and 27 of 29 (Appendix D). The failure modes this set exposed are listed in Appendix H and left unfixed, so that the set stays what it is. Instrument error, not sampling, is the larger uncertainty on the failure rate, and it is propagated as recall was: the verdict disagreement of 2 of 68 has upper bound 10.1%, and if that share of either verdict were wrong the headline would sit in 42.8–52.9% (Table 1); the observed disagreements all run one way, the pipeline surviving where the reader fails, which puts it at 49.1%, or 55.7% dropping the 32 items the reader could not test (15 not a single-bold claim to it, 17 without a usable run count; Appendix D). The claim disagreement of a third is mostly in what the pipeline compares, not in the values it reads.

**The headline without the pipeline.** On 247 table pages drawn at random from the corpus, one per paper, the same reader listed every results table and single-bold group with nothing marked and no pipeline output in view (Appendix D). From its labels alone, 80.7% [70.7, 88.9] of 1657 groups carry no error bar on either side and 12.3% [6.0, 19.6] are testable; the pipeline's comparisons on the same pages give 87.6% and 9.7% of 1076; the draw carries fewer bars than the corpus (weighted one page per paper, the corpus gives 79.0%), not a different weighting. The instruments disagree on the census and, less, on the shares: the reader lists 77.9% of the pipeline's groups, the pipeline only 50.6% of the reader's; the groups it misses print bars more often (76.3% bar-less against 85.0%), so over every group the reader finds the share is 0.921 times the pipeline's (72.4% for the headline). From its own labels the reader fails 30.7% [17.8, 46.2] of the 192 groups it can test, the pipeline 59.6% of its 104 on the same pages, and most of the difference is in which groups are testable: the two agree on 61 of the 66 groups both can test (26 fail by the reader's values, 31 by the pipeline's) and part company where they cannot: the reader tests 126 groups the pipeline does not and fails 33; the pipeline tests 38 the reader does not, on a run count read from elsewhere in the paper, and fails 31. Which run counts one accepts decides the failure rate as much as the test does, which is the caption-or-page row of Table 2 seen from the other side.

The weakest field is the run count, read from the paper rather than the table. On the held-out items the reader found none in 20 of 49; where it found one, the pipeline's agreed in 27 of 29. A second reading with the whole paper adjudicated the 14 of those 20 to which the released pipeline attaches a count: 11 were the first reader's misses, 3 (one paper) an oracle's trial count quoted in a prompt. Where the pipeline attaches a run count it is the paper's in 38 of 43 (88.4% [75.5, 94.9]); Table 2 scales the checkable share by that precision, and restores the counts a third draw of 40 bar-carrying comparisons found stated in words the pattern had missed (18 of 40).

**Recall.** The instrument reads table cells and nothing else: uncertainty in a figure, an image table or shading is invisible to it and counted as absent (Appendix K counts what lies outside the cells, Appendix A tie-marked tables, Appendix C the 411 NeurIPS checklists reporting error bars we cannot find). Precision does not bound the quantity the headline rests on, which is recall: an error bar present in the PDF and missed falls silently into "neither value has an error bar". We rendered 149 tables the extractor called empty and looked for any error bar at all. The audit found four misses, three of them bars written without a leading zero ($0.444 \pm .003$) and one a table whose winning column the block-cut heuristic had dropped; both are fixed (Appendix H). The recall nothing was fixed on is the pre-fix one, 97.3% [93.3, 99.0]. Under the released pipeline 4 of the 149 now read a bar, 14 yield no claim any more, and of the 131 still called empty it misses none (recall 95% CI [0.97, 1.00], on the tables the fixes came from). Both are conditional on a table being detected and its bars being in its cells: the comparison-level audit of the same stratum on the held-out draw found 95.0% [83.5, 98.6] truly bar-less (39 of 40 on the frozen draw), the exceptions being dispersions printed in a separate column, which the cell grammar cannot pair with a mean. Scaled by that rate the share carrying no error bar is 65.6–77.5%, beside the 76.4–78.6% from the table-level audit. We quote the propagated intervals beside the sampling interval 78.6% [75.1, 82.0], so that the uncertainty of our own headline is reported as well. A check that depends on neither the PDF nor the reader, the arXiv LaTeX source of 217 papers, agrees with the pipeline on whether any table prints an error bar in 91.2% [86.7, 94.3] of papers and finds one in 82.4% [72.2, 89.4] of the papers whose source prints one (Appendix E).

**The testable subset may be unrepresentative.** The 5465 testable comparisons come from 245 papers that report uncertainty, and whether their failure rate sits above or below the rate every paper would show is not established. Three cuts say the selection is not large: reweighting to the population's subfield mix gives 45.4% and to its distribution of relative gap, the one quantity observed for every comparison, 46.5%; ablation tables (14.0% of comparisons) fail at 54.7% [42.3, 67.0] against 46.5% [41.3, 51.6] for the rest; and by stated run count the rate is 56.5% [50.7, 62.6] at three or fewer, 51.4% [43.1, 59.3] at four or five and 36.6% [26.4, 46.1] at six or more. We claim no bound.

## 10 Related work

Appendix J discusses each line of prior work in detail.

**Reporting, re-running, and hand audits.** One line of work tells authors what they should print (Dodge et al., 2019; Bouthillier et al., 2021; Miller, 2024; Bowyer et al., 2025); another re-runs a literature and finds the gains smaller than claimed (Henderson et al., 2017; Musgrave et al., 2020; Dacrema et al., 2019; Saidi et al., 2024); we measure how often any interval is printed at all. Table 7 sets the closest hand audits beside this one (Berg-Kirkpatrick et al., 2012; Dror et al., 2018; Gundersen & Kjensmo, 2018; Király et al., 2018; Pham et al., 2020; Hutchinson et al., 2022; Christodoulou et al., 2024; 2025). Their figures are not ours on a larger corpus: an error bar anywhere in a paper, several runs, a significance test, and an interval in a results table are different events, ours the narrowest, so our paper-level 26.0% sits below the earlier figures. Working at the comparison yields 32300 observations where hand annotation yields hundreds, at the cost of judgement on an ambiguous table, which Section 9 measures. Pineau et al. (2021) describe the NeurIPS checklist; Magnusson et al. (2023) and Goldberg et al. (2024) compare its answers with paper content, as Appendix C does.

**Statistical practice as an object of study.** Card et al. (2020), Colas et al. (2018) and Lin et al. (2021) compute the power of experiments or the significance of leaderboard gaps in one setting each; Section 7 asks that question by machine of every named-benchmark comparison in two venues, on the benchmark variance Madaan et al. (2024) measure. Reimers & Gurevych (2017), Agarwal et al. (2021), Ulmer et al. (2022), Marie et al. (2021) and Tang et al. (2020) study one subfield's practice by hand; Nuijten et al. (2016) recompute the $p$-values printed in psychology papers. Hou et al. (2019), Kardas et al. (2020) and Park et al. (2025) extract result tables at scale to populate leaderboards, the same mechanics with the opposite purpose; Bianchi et al. (2025) audit papers for errors with a language model whose flags were verified by hand, the human half Section 9 lacks. No released comparison-level annotation of error-bar presence exists; the nearest, the paper-level confidence-interval label of Király et al. (2018), counts an interval anywhere in a paper, so Section 9 builds one.

## 11 Discussion

The cheapest fix needs no new statistics, only three quantities authors already have: a dispersion on each value in the main comparison and the number of runs behind them. Printed, they make the comparison checkable by any reader at no cost. A second measure is to consult Appendix B while the experiment is being designed, where an author will sometimes find that a test set of that size cannot resolve the intended claim, and can then say so or change the design.

### Reproducibility statement

The corpus is defined by a venue, a year, and a random seed, and is rebuilt by one command from the open proceedings; a second command regenerates every analysis, figure and number from the extraction files, deterministically. Every number in this paper is a macro generated from a recorded run; a number typed into the LaTeX by hand fails the build. The extraction pipeline, the analysis, the per-comparison database, the reference-set renderings with their labels and scoring script (including the frozen set and the tag of the pipeline it was drawn under), and a regression test for every extractor bug described in this paper are released at `https://anonymous.4open.science/r/checkable--E2C5`.

### AI USE STATEMENT

The authors used a large language model assistant throughout: to write and revise the extraction and analysis code, to draft and edit prose, and to search the literature. Every empirical number in the paper was produced by the released code and checked against its run record. The rendered-table audits of Section 9 (the recall audit, the rejected-table audit and the reference set) were read from the images by the assistant (Claude Opus 5, reading PNG renderings, with the prompt given in Appendix D; the other readers of Appendix D are Claude Sonnet 5 and Gemini 3.1 Pro), with the pipeline's output hidden from them; the renderings and labels are released so that any reader can re-score them, and the authors are responsible for the protocol, the labels and every claim in the paper.

### ETHICS STATEMENT

This work audits published papers in aggregate. We report no individual paper as an example of a failure and name no authors. The per-comparison database is released with paper identifiers so that our claims can be checked, and we say plainly that sorting it by test statistic identifies papers; we release it anyway, because a checkability audit that could not itself be checked would be its own counterexample. The pattern is a property of the field's conventions, not of the people following them.

## REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2108.13264.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

Federico Bianchi, Yongchan Kwon, Zachary Izzo, Linjun Zhang, and James Zou. To err is human: Systematic quantification of errors in published AI papers via LLM analysis. 2025. URL https://arxiv.org/abs/2512.05925.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. 2021. URL https://arxiv.org/abs/2103.03098.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. 2025. URL https://arxiv.org/abs/2503.01747.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. 2020. URL https://arxiv.org/abs/2010.06595.

Evangelia Christodoulou, Annika Reinke, Rola Houhou, Piotr Kalinowski, Selen Erkan, Carole H. Sudre, Ninon Burgos, Sofiène Boutaj, Sophie Loizillon, Maëlys Solal, Nicola Rieke, Veronika Cheplygina, Michela Antonelli, Leon D. Mayer, Minu D. Tizabi, M. Jorge Cardoso, Amber Simpson, Paul F. Jäger, Annette Kopp-Schneider, Gaël Varoquaux, Olivier Colliot, and Lena Maier-Hein. Confidence intervals uncovered: Are we ready for real-world medical imaging AI? 2024. URL https://arxiv.org/abs/2409.17763.

Evangelia Christodoulou, Annika Reinke, Pascaline Andrè, Patrick Godau, Piotr Kalinowski, Rola Houhou, Selen Erkan, Carole H. Sudre, Ninon Burgos, Sofiène Boutaj, Sophie Loizillon, Maëlys Solal, Veronika Cheplygina, Charles Heitz, Michal Kozubek, Michela Antonelli, Nicola Rieke, Antoine Gilson, Leon D. Mayer, Minu D. Tizabi, M. Jorge Cardoso, Amber Simpson, Annette Kopp-Schneider, Gaël Varoquaux, Olivier Colliot, and Lena Maier-Hein. False

promises in medical imaging AI? assessing validity of outperformance claims. 2025. URL https://arxiv.org/abs/2505.04720.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. 2018. URL https://arxiv.org/abs/1806.08295.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys 2019)*, 2019. doi: 10.1145/3298689.3347058. URL https://arxiv.org/abs/1907.06902.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. 2019. URL https://arxiv.org/abs/1909.03004.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. doi: 10.18653/v1/P18-1128.

Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment. 2024. URL https://arxiv.org/abs/2411.03417.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11503.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. 2017. URL https://arxiv.org/abs/1709.06560.

John M. Hoenig and Dennis M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001. doi: 10.1198/000313001300339897.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/P19-1513.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. 2022. URL https://arxiv.org/abs/2205.05256.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. AxCell: Automatic extraction of results from machine learning papers. 2020. URL https://arxiv.org/abs/2004.14356.

Franz J Király, Bilal Mateen, and Raphael Sonabend. NIPS - not even wrong? a systematic review of empirically complete demonstrations of algorithmic effectiveness in the machine learning and artificial intelligence literature. 2018. URL https://arxiv.org/abs/1812.07519.

Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? a case study of the MS MARCO document ranking leaderboard. 2021. URL https://arxiv.org/abs/2102.12887.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. 2024. URL https://arxiv.org/abs/2406.10229.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. Reproducibility in NLP: What have we learned from the checklist? 2023. URL https://arxiv.org/abs/2306.09562.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 2021. URL https://arxiv.org/abs/2106.15195.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. 2024. URL https://arxiv.org/abs/2411.00640.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. 2020. URL https://arxiv.org/abs/2003.08505.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48:1205–1226, 2016. doi: 10.3758/s13428-015-0664-2.

Jungsoo Park, Junmo Kang, Gabriel Stanovsky, and Alan Ritter. Can LLMs help uncover insights about LLMs? a large-scale, evolving literature analysis of frontier LLMs. 2025. URL https://arxiv.org/abs/2502.18791.

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE 2020)*, 2020. doi: 10.1145/3324884.3416545.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 2021. URL https://arxiv.org/abs/2003.12206.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. doi: 10.18653/v1/D17-1035.

Pouria Saidi, Gautam Dasarathy, and Visar Berisha. Unraveling overoptimism and publication bias in ML-driven science. 2024. URL https://arxiv.org/abs/2405.14422.

Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin. Showing your work doesn't always work. 2020. URL https://arxiv.org/abs/2004.13705.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. URL https://arxiv.org/abs/2204.06251.

## A TIE MARKING

A paper that bolds the best result and marks every result not statistically distinguishable from it has already declared those pairs to be ties; counting them as failed comparisons would inflate our failure rate. The convention is real but rare, and it concentrates where it matters. Captions claiming statistical indistinguishability appear in 0.12% of all tables but 0.24% of the testable subset, because the papers careful enough to print dispersions are the ones most likely to mark ties. Those comparisons fail at 84.6% against 47.6% overall, which is what one would expect of pairs the authors had already declared indistinguishable.

The mechanism is therefore confirmed and the magnitude is small. Removing tie-marked comparisons moves the failure rate from 47.6% to 47.5%, because the convention remains rare even after the enrichment. The commoner convention, marking the second-best result (4.1% of tables), is not a tie claim and is left in.

Table 3: Smallest difference in accuracy, in percentage points, that each benchmark can distinguish from zero at 80 per cent power. Columns give the independent reading at three accuracies and the paired reading at 70 per cent. A property of the benchmark, not of any paper.

| Benchmark | Split | At 50% | At 70% | At 90% | Paired, 70% |
|---|---|---|---|---|---|
| AIME | 30 | 36.17 | 33.15 | 21.70 | 23.44 |
| HumanEval | 164 | 15.47 | 14.18 | 9.28 | 10.03 |
| GPQA Diamond | 198 | 14.08 | 12.90 | 8.45 | 9.12 |
| MBPP | 500 | 8.86 | 8.12 | 5.32 | 5.74 |
| OpenBookQA | 500 | 8.86 | 8.12 | 5.32 | 5.74 |
| MATH-500 | 500 | 8.86 | 8.12 | 5.32 | 5.74 |
| SWE-bench Verified | 500 | 8.86 | 8.12 | 5.32 | 5.74 |
| TruthfulQA | 817 | 6.93 | 6.35 | 4.16 | 4.49 |
| SST-2 | 872 | 6.71 | 6.15 | 4.03 | 4.35 |
| ARC-Challenge | 1,172 | 5.79 | 5.30 | 3.47 | 3.75 |
| WinoGrande | 1,267 | 5.57 | 5.10 | 3.34 | 3.61 |
| GSM8K | 1,319 | 5.45 | 5.00 | 3.27 | 3.54 |
| PIQA | 1,838 | 4.62 | 4.24 | 2.77 | 2.99 |
| ARC-Easy | 2,376 | 4.06 | 3.72 | 2.44 | 2.63 |
| BoolQ | 3,270 | 3.46 | 3.18 | 2.08 | 2.25 |
| MATH | 5,000 | 2.80 | 2.57 | 1.68 | 1.82 |
| QNLI | 5,463 | 2.68 | 2.46 | 1.61 | 1.74 |
| CoQA | 7,983 | 2.22 | 2.03 | 1.33 | 1.44 |
| DROP | 9,536 | 2.03 | 1.86 | 1.22 | 1.31 |
| MNLI | 9,815 | 2.00 | 1.83 | 1.20 | 1.30 |
| CIFAR-10 | 10,000 | 1.98 | 1.82 | 1.19 | 1.28 |
| CIFAR-100 | 10,000 | 1.98 | 1.82 | 1.19 | 1.28 |
| HellaSwag | 10,042 | 1.98 | 1.81 | 1.19 | 1.28 |
| SQuAD | 10,570 | 1.93 | 1.77 | 1.16 | 1.25 |
| MMLU-Pro | 12,032 | 1.81 | 1.66 | 1.08 | 1.17 |
| MMLU | 14,042 | 1.67 | 1.53 | 1.00 | 1.08 |
| ImageNet | 50,000 | 0.89 | 0.81 | 0.53 | 0.57 |

## B What each benchmark can resolve

Table 3 gives, for each benchmark in common use, the smallest difference in accuracy it can distinguish from zero at $\alpha = 0.05$ and 80% power. The quantity depends only on the published evaluation split, so it is fixed, checkable, and available before an experiment is run.

Two readings are shown. The independent-binomial column treats the two systems as unrelated, which is conservative; the paired column assumes they are scored on the same items with a correlation of 0.5, which is closer to how benchmarks are actually used and resolves finer. Values are given at three accuracies because a binomial variance is largest near one half, so a benchmark resolves more finely at the top of its scale than in the middle.

## C The checklist and the recall audit

The recall audit says the extractor misses almost nothing; the checklist says 411 NeurIPS papers claim error bars we cannot find. Both cannot be right unless the claim means something other than what we measure, so we read what those justifications actually say.

Magnusson et al. (2023) and Goldberg et al. (2024) read checklist answers against paper content for NLP and for NeurIPS 2024 respectively; ours is the same comparison for one question. The justifications mostly do not claim an error bar in a results table. The commonest, at 30.4%, is that several random seeds were used, which averages the runs without reporting their spread. Another 18.7% report having run a significance test, 13.9% point to a figure and 14.1% to an appendix, and 30.4% make no recognisable claim at all (the categories overlap, and the first and last shares coincide by chance). The question allows "other appropriate information about the statistical significance", so the papers citing a test or a figure may well be compliant; our instrument does not read either, and the contribution is stated accordingly: 411 papers print no error bar in any results table, not

Table 4: Reported gaps against what the benchmark can resolve, for every benchmark with thirty or more matched comparisons, treating the two systems as independent. Gap and resolution (the minimum detectable difference at 80% power) are medians in points at the comparison's own pooled accuracy; 'Not significant' is the share of gaps below 1.96 standard errors of the difference, 'Below resolution' the share below the 80%-power value. Neither is a function of split size alone: both depend on the gaps a field reports on that benchmark.

| Benchmark | Split | $n$ | Gap | Resolution | Not significant | Below resolution |
|---|---|---|---|---|---|---|
| ImageNet | 50,000 | 610 | 1.2 | 0.8 | 25.7% | 34.3% |
| MMLU | 14,042 | 58 | 1.4 | 1.5 | 34.5% | 43.1% |
| CIFAR-10 | 10,000 | 497 | 1.4 | 1.5 | 40.4% | 49.7% |
| CIFAR-100 | 10,000 | 283 | 1.4 | 1.7 | 44.9% | 55.5% |
| DROP | 9,536 | 68 | 1.8 | 1.0 | 27.9% | 30.9% |
| MATH | 5,000 | 343 | 0.5 | 2.6 | 77.0% | 80.8% |
| GSM8K | 1,319 | 79 | 1.6 | 5.2 | 74.7% | 79.7% |
| WinoGrande | 1,267 | 79 | 0.5 | 5.1 | 96.2% | 98.7% |
| SST-2 | 872 | 38 | 6.7 | 6.4 | 28.9% | 50.0% |
| TruthfulQA | 817 | 43 | 2.3 | 6.2 | 72.1% | 79.1% |
| AIME | 30 | 38 | 40.8 | 29.8 | 36.8% | 36.8% |

411 papers with nothing to show. The two measurements are consistent, and the disagreement is the finding: what the checklist records is not what it asks.

Two extraction limits could have produced the same gap and do not. Papers longer than our 40-page window are 12.9% of the corpus and have a *lower* error-bar rate (24.1% against 26.2%), so truncation is not hiding a mass of error bars in appendices. And 12.8% of papers yield no detected table at all; excluding them raises the paper-level rate from 26.0% to 29.8%, and we report both. The paper-level statistic reads every detected table, not only those reaching the comparison stage: of the 471 papers whose tables yield no comparison, 95 are correctly credited with an error bar. Only papers with no detected table are scored blind.

## D THE REFERENCE SET

Table 5: The reference set. Each item is rendered from the PDF with the pipeline's chosen cell boxed and read blind to its output; a digit the reader could not make out was recorded as unreadable and left out of that field. The development set found the bugs of Appendix H; the held-out set, drawn from papers the development set never touched, found two more and is scored before and after them; the frozen set, drawn after the released pipeline was tagged and never used to fix anything, is the estimate the paper stands on.

| Stratum | Drawn from | $n$ | What is read |
|---|---|---|---|
| Development | testable comparisons, pre-fix pipeline | 100 | every field |
| Held-out | testable comparisons, fixed pipeline | 50 | every field |
| Frozen | testable comparisons, released pipeline | 100 | every field; is the rival copied |
| Frozen, no bar | retained comparisons without bars | 40 | is a bar really absent |
| Frozen, no run count | both bars, no run count found | 40 | does the whole paper state one |
| Frozen, unboxed | the frozen crops, nothing marked | 100 | every single-bold group |
| Bar-less, count attached | no bars, count read from the paper | 40 | does the count apply here |
| Random pages | a table page per paper, nothing marked | 247 | every table and group on it |
| No run count, third draw | both bars, no run count found | 40 | does the paper state one |
| No error bar, held-out draw | retained comparisons without bars | 40 | is a bar really absent |
| Discarded | tables with no single-bold group | 40 | why it was discarded |
| Non-extremal | single-bold groups, bold not best | 40 | what the bold marks |

**What is thrown away.** The audit also read 40 tables the pipeline discards for lacking a single-bold group and 40 groups whose bold cell is not the best. Of the discarded tables, 40.0% mark no winner at all, 32.5% are not results tables, 20.0% mark winners in a way the pipeline does not

read (shading, colour, several per column), and 7.5% are the pipeline's fault. Of the non-extremal groups, 2.5% are authors marking their own row regardless of rank, 70.0% mark the best within a sub-block or excluding an oracle, 12.5% are misreads, and 15.0% mark something else (a selected hyperparameter, a column mixing units). Of 40 retained comparisons the pipeline says carry no error bar on either side, 95.0% [83.5, 98.6] carry none; the exceptions are tables whose dispersion sits in a separate column, which the instrument cannot pair with a mean.

**Protocol.** Each item is a crop of the page around the table at twice its printed size, with a red box drawn around the cell the pipeline took as the winner and nothing else drawn; the reader also receives every sentence of the paper containing a word such as seeds, runs, trials or times, so that it sees the same evidence for a run count the pipeline does. The reader is asked, per item: is this a results table; is the boxed cell bold and alone in its group; is the group a row or a column; is higher or lower better; the boxed cell's mean and error bar and those of the best unbolded cell in the group, as printed; whether the caption or paper says the bar is a deviation, an error, an interval, or nothing; and the run count, with where it was found. A digit the reader cannot make out is recorded as unreadable and left out of that field's denominator. An earlier rendering drew a circle over the cell, which covered the digits in a third of the items; those were re-rendered with the box and re-read. Every field is scored against the pipeline's value by exact match, and a verdict is recomputed from the reader's values with the paper's own test. The reader's instruction, verbatim: "For this item record: results_table (y/n); bold_ok (y/n: the boxed cell is bold and the only bold cell in its group); orientation (row/column); direction (higher/lower); winner_mean, winner_disp, runner_mean, runner_disp as printed, 'none' if absent, 'unclear' if you cannot read a digit; disp_type (sd/se/ci/unlabelled) from the caption or the paper; n_runs and n_source (caption/page/paper/none) from the sentences supplied; a note." The frozen set added one question, whether the runner-up's row or column carries a citation, dagger or "from" marker, and for its no-run-count stratum gave the reader the whole paper rather than the selected sentences. Its items were drawn by simple random sampling of comparisons, so several land in one paper; the claim agreement is therefore also given with an interval that resamples papers, and the per-field agreement on it is: bold cell 85.0%, orientation 88.0%, direction 96.0%, winner 100.0% and its bar 91.9%, runner-up 78.8% and its bar 72.7%, bar kind at the level the test uses 88.0% (on the held-out set 92.0%, and on the exact label only 54.0%), run count 83.0%. In the unboxed control the reader was asked, for each crop, whether it is a results table, which grouping the table marks winners within, how many cells are bold, and for every group containing exactly one bold cell its header, the direction, the bold cell's mean and bar and the best non-bold cell's mean and bar; a group is matched to a pipeline comparison by the winner's printed value. The reader found no single-bold group at all in 12 of the 100 crops, most of them the one paper whose bold marks every valid method. Of the 32 boxed items the reader could not test, the pipeline fails 3 of the 15 it judged not a single-bold claim (5 papers) and 6 of the 17 to whose bars it could attach no run count (10 papers). The bar-less stratum was read as the no-run-count one, with the whole paper, the reader stating the count that applies to the boxed table or none; in 18 of 40 the paper states no count for that table and in 1 of 40 a different one. For the random pages, 256 papers were drawn and one page containing a table caption chosen at random in each; the reader received the whole page and the paper's run sentences, and was asked for every table of experimental results on it, its grouping, whether any cell prints a dispersion, the bar kind named in the caption, the run count that applies, and every single-bold group with the same fields as the boxed items. The reader's content filter declined 9 pages, which are excluded; of 292 results tables on the rest, 22.9% print any dispersion. Groups are matched to pipeline comparisons on the same page by the winner's printed value; intervals resample pages.

**Other readers.** The sixty-item labelling kit released with the frozen set (the first forty testable items and the first twenty no-bar items) was also read blind, under the same prompt, by a smaller model of the same family (Claude Sonnet 5) and by a model from another vendor (Gemini 3.1 Pro). Against the first reader the two agree on the bold cell in 53 of 58 and 51 of 58 of the items both could judge, the winner in 52 of 58 and 57 of 58, its bar in 35 of 39 and 38 of 39, the runner-up in 54 of 58 and 54 of 57, its bar in 36 of 39 and 37 of 39, and the run count in 37 of 40 and 35 of 40. Against the pipeline they agree on the whole claim in 26 of 40 and 28 of 39 and on the verdict in 28 of 29 and 27 of 29, and both find no bar on every no-bar item (20 of 20, 20 of 20). Three models of two vendors agreeing narrows the shared blind spot but is not a human reference. One author made a single pass over the 60 items with the pipeline's reading shown and one key per item, in 3.9 minutes, marking 12 correct, 25 wrong and 23 unclear, and reported no confidence in the result before it was scored;

its judgements agree with the majority of the three machine readers on 16 of 37 items (43.2%), near chance. The pass is released as a record and no number in this paper rests on it; the kit is released for a reader with time to look.

**Every scoring.** Table 6 prints every comparison of the reader's verdicts with the pipeline's, before and after the fixes each set prompted, with the number of items both sides could test. On the development and frozen sets the pipeline calls fewer comparisons non-significant than the reader does, because the reader attaches run counts and bar meanings stated elsewhere in the paper; on the two held-out rows the difference is one item the other way.

Table 6: Reader against pipeline on the reference sets. "Claim" is agreement on the bold cell, the rival and all four values on fully readable items; "verdict" is per-item agreement on fail or survive among items both sides could test; the last two columns are the failure rate on those items from each side's values.

| Set | Pipeline | Claim | Verdict | Reader fails | Pipeline fails |
|---|---|---|---|---|---|
| Development | before its fixes | 87 of 97 | 70 of 82 | 52.4% | 42.7% |
| Development | released | 89 of 94 | 72 of 81 | 51.9% | 48.1% |
| Held-out | before its fixes | 40 of 50 | 31 of 32 | 50.0% | 53.1% |
| Held-out | released | 36 of 41 | 28 of 29 | 48.3% | 51.7% |
| Frozen | released, nothing fixed | 69 of 99 | 66 of 68 | 54.4% | 51.5% |

## E THE LATEX-SOURCE CROSS-CHECK

We matched 300 papers of the 2025 sample to arXiv by title, downloaded the source of the 224 that matched, and read `\pm` and `\textbf` inside every table environment of the 217 that have one. At the paper level the source and the pipeline agree on whether any table prints an error bar in 91.2% [86.7, 94.3]; of the 74 papers whose source prints one, the pipeline finds it in 82.4% [72.2, 89.4], the rest being tables the pipeline did not detect or an appendix beyond its page window. Matching tables by caption gives 935 pairs, on which the two agree about the presence of an error bar in 92.7% [90.9, 94.2] (21 tables where the source has one and the PDF read none, 47 the reverse) and about the presence of a bold number in 86.4% [84.1, 88.5]. Of the papers whose source prints a bar and the pipeline reads none, 3 have that table matched to a PDF table read as empty, 6 have it undetected in a paper within the page window, and 4 are papers longer than the window.

The script `latex_crosscheck.py` queries the arXiv API with each paper's title, accepts a match whose title agrees at similarity 0.9 or better, downloads the e-print, inlines `\input` files, strips comments, and reads every `\begin{table}` environment: its caption, whether the tabular body contains `\pm` (or `\mp`, or the glyph), and whether it contains `\textbf` or `\mathbf` applied to a number, which is the event the pipeline reads; a header set in bold is not. Source tables and PDF tables are paired by caption similarity at 0.6 or better. Disagreement counts every table that differs between the arXiv and proceedings versions, every table the PDF pipeline failed to anchor on a caption, and every macro that hides a `\pm` behind a name, so it is an upper bound on the extractor's error. Of the 31 tables where the source bolds a number and the PDF read no bold cell, most are the `\mathbf` and colour-coded conventions Section 9 already counts as invisible; 96 run the other way. The cached source tables are released with the audits.

## F PRIOR AUDITS

Table 7 sets the hand audits of reporting practice beside this one. Their headline figures are not comparable to ours: an error bar anywhere in a paper, a confidence interval, a significance test, and an explicit interval in a results table are four different events, and ours is the narrowest.

## G BY SUBFIELD

Table 8 assigns each paper a coarse subfield from its title by keyword, in the order listed, and reports both headline figures within each. The assignment is crude and is meant only to answer whether the

Table 7: Audits of reporting practice. Each earlier study labels papers by hand, at the scale of a person; this one labels comparisons by machine, and Section 9 measures the instrument against rendered tables read by a second, image-based reader and against LaTeX source.

| Study | Venue | By | Unit | Headline |
|---|---|---|---|---|
| Berg-Kirkpatrick et al. (2012) | ACL | hand | paper | significance tests rarely reported |
| Dror et al. (2018) | ACL, TACL 2017 | hand | paper | about a third test significance |
| Gundersen & Kjensmo (2018) | AAAI, IJCAI | hand | paper | most omit what a reproduction needs |
| Király et al. (2018) | NeurIPS 2017 | hand | paper | about a third report intervals |
| Pham et al. (2020) | AI, SE, systems venues | hand | paper | a fifth use several runs |
| Hutchinson et al. (2022) | NLP, vision conferences | hand | paper | error bars typically not reported |
| Christodoulou et al. (2024) | MICCAI 2023 | hand | paper | over half assess no variability |
| Christodoulou et al. (2025) | MICCAI | hand | claim | most outperformance claims at risk |
| This paper | NeurIPS, ICML 2024, 2025 | machine | comparison | 78.6% print no error bar |

Table 8: Both headline figures by coarse subfield, assigned from the title by keyword in the order listed. The share without an error bar is a property of the subfield; the failure rate among what can be tested is not.

| Subfield | Papers | Comparisons | No error bar | Testable | $n$ | Not distinguishable |
|---|---|---|---|---|---|---|
| Reinforcement learning | 183 | 2450 | 58.0% | 34.9% | 856 | 46.3% |
| Graphs and molecules | 27 | 489 | 49.1% | 44.6% | 218 | 48.2% |
| Vision and generation | 406 | 8512 | 88.7% | 8.6% | 736 | 37.1% |
| Language models | 247 | 5773 | 89.6% | 7.6% | 441 | 42.6% |
| Theory and optimisation | 143 | 996 | 54.9% | 23.2% | 231 | 48.5% |
| Other | 792 | 14080 | 74.3% | 21.2% | 2983 | 51.1% |

headline is the property of one community. It is not: reporting is worst where evaluation is cheapest, in language-model and vision benchmarks, and the failure rate among what can be tested is similar everywhere there is enough to test. The share of gaps below the imputed noise of Section 6 runs from 37.8% in language-model papers to 56.9% in reinforcement learning.

## H Every extractor bug, and what it did

All sixteen were found by rendering a sampled table and looking, or by the LaTeX-source cross-check; none by reading code. Each has a regression test in the released code.

1. **Subscript error bars** set a point below the baseline were paired with the neighbouring column's mean.
2. **The page number** beneath a table was swept in as a runner-up.
3. **The neighbouring column** of a two-column layout was swept in.
4. **Orientation misread**: a table with methods in columns was read as methods in rows, comparing a method with itself on another benchmark.
5. **A run count from a row label** reading "top 100".
6. **Bracketed bars** $76.63(\pm 3.98)$ fell between two patterns and were read as bare numbers; correcting it moved the headline by 0.5 points.
7. **Rows glued by a tall label**: a label spanning two six-point rows widened the line so the next row joined it, and $0.67$ beside $0.00 \pm 0.00$ became $670.0 \pm 0.0$.

8. **A row set as one span** with no space characters read as $0.7015 \pm 0.0007 0.2577 \pm 0.0046$ and yielded a winner of $2577$.
9. **Dataset blocks**: one bold cell per block made every column multi-bold, the orientation vote flipped to rows, and a method's calibration error was compared with its expected calibration error.
10. **Bare-integer row labels** ("# demos $= 10$") outranked the bold cell, so the claim vanished as non-extremal.
11. **Run counts in words** ("four seeds", "ten runs") were invisible, and a task named "Run" beneath a cell ending in 95 gave $n = 95$.
12. **Bars without a leading zero** ($0.444 \pm .003$) were read as bare numbers; a whole table of them was called empty.
13. **Bold weights without the bold flag**, whose font name says "Medi" or "CMBX" rather than "bold", were read as regular, and a table whose winning row was set that way was discarded.
14. **The window above a caption**, kept for venues that caption underneath, reached into the table before it and swept in that table's last rows, whose bold cells outranked the real winner.
15. **Run counts in other phrasings**: "each experiment is repeated five times", "three randomly seeded runs", and a list of the seeds themselves were invisible to the pattern, and the comparison was called untestable.
16. **A caption below its table** read as a figure: after fix fourteen the window above a caption was searched first, and a table whose caption sits underneath had its rows read as the axis ticks of the figure above; the extractor now reads both windows and keeps the one with grid-shaped rows.

Bugs seven to eleven were found by the development reference set, twelve by the recall audit, thirteen and fourteen by the held-out reference set, fifteen by the run-count audit (Section 9), and sixteen by the LaTeX-source cross-check (Appendix E). Together they moved the 2025 headline from 77.5% to 78.6% comparisons without an error bar and the failure rate from 40.6% to 47.6%; every intermediate version of these numbers was, at the time, confidently reported.

**Found by the frozen set and not fixed.** The frozen set of Section 9 exposed further failure modes, left in the released pipeline so that the set remains an estimate nothing was fitted to: a table whose bold marks every valid method rather than a winner, with two metrics per row, read with rows as groups (one paper, 11 items); a two-line cell whose second line is another metric, paired as the rival; a negative sign dropped from a value; a cell written as "before $\rightarrow$ after $\pm$ sd"; a coloured second-best mark read as bold; and a table in which every mean is bold. Each is a candidate for the next version, and each will be measured on a set drawn after it.

## I Secondary sensitivity figures

**The binomial range filter.** Section 7 restricts to comparisons whose two values both lie in $[40, 100]$. Widening the filter to all values in $(0, 100]$ raises the median test-set requirement from 10,097 to 30,607 items, because it admits smaller gaps; the share needing more than GSM8K moves by a point.

**Unlabelled bars.** On the 1929 unlabelled comparisons with a run count, reading every bar as a standard deviation gives 47.2% failing and as a standard error 69.1%; carried into the full population of 5465 testable comparisons the second reading gives 55.3% (Table 2). The headline uses the first, the more conservative, and the subset whose authors state what the bar means fails at 47.8%.

## J Related work in detail

**How to report uncertainty.** A line of work tells authors what they should print. Dodge et al. (2019) argue for reporting the full distribution of results over hyperparameter search rather than a

single tuned number; Bouthillier et al. (2021) decompose the variance of a benchmark result into its sources and show how many seeds are needed to detect a given effect; Miller (2024) derives the error bars a language model evaluation should carry, and Bowyer et al. (2025) shows that the usual normal approximation fails on the small evaluation sets now common. These papers say what a correct interval looks like. We measure how often any interval is printed at all, which is prior to whether it is correct.

**Whether reported gains are real.** A second line re-runs a literature under a common protocol and finds the gains smaller than claimed or absent: Henderson et al. (2017) in deep reinforcement learning, Musgrave et al. (2020) in metric learning, Dacrema et al. (2019) in neural recommendation. Saidi et al. (2024) find the same pattern across machine-learning-driven science and attribute it partly to publication bias. Each of these requires re-running the methods, which bounds them to one subfield at a time. Our measurement needs only the printed page, so it covers whole venues, at the cost of saying nothing about whether any individual result is true.

**Hand audits of reporting practice.** A line of studies asks our question by reading papers one at a time, and Table 7 sets them beside this one. Berg-Kirkpatrick et al. (2012) and Dror et al. (2018) count significance testing in computational linguistics; Gundersen & Kjensmo (2018) score AAAI and IJCAI papers for what a reproduction would need; Király et al. (2018) review NeurIPS 2017 and find roughly a third reporting confidence intervals; Pham et al. (2020) find about a fifth of papers at AI, software-engineering and systems venues training more than once; Hutchinson et al. (2022) hand-code papers from recent NLP and computer-vision conferences and find error bars typically not reported; Christodoulou et al. (2024) annotate 221 MICCAI 2023 papers and find more than half assessing no variability at all, and Christodoulou et al. (2025) estimate, per claim, the probability that an outperformance claim in the same literature arose by chance; Agarwal et al. (2021) show that point estimates over a handful of runs misorder deep reinforcement learning methods. Pineau et al. (2021) describe the NeurIPS checklist we read in Section 9. Their figures are not ours on a larger corpus, because the definitions differ: an error bar anywhere in a paper, a confidence interval, a significance test, and an explicit interval in a results table are four different events, and ours is the narrowest. Our paper-level 26.0% is accordingly below the earlier figures, not between them.

Their limit is the one manual annotation has: a few hundred papers, and a label per paper. The unit a results table claims in is the comparison, and a paper printing one error bar in an appendix while its main table prints none counts as reporting variability under a paper-level audit and as unchecked under ours. Working at the comparison is what makes 32300 observations available where hand annotation yields hundreds, and what makes Sections 5 to 7 askable: whether the claimed difference survives its own error bars, whether a design of that size could have detected it, and whether the benchmark can resolve it. What automation costs is judgement on an ambiguous table, which is why Section 9 measures what ours gets wrong.

**Statistical practice as an object of study.** A second group asks the question of one subfield each. Card et al. (2020) compute the statistical power of natural language processing experiments and find most underpowered, and Colas et al. (2018) do the same for deep reinforcement learning; the comparison of a claimed gap against what a benchmark can resolve (Section 7) is theirs, carried out here by machine on every named-benchmark comparison in two venues. Reimers & Gurevych (2017) show that a single run can misorder methods, and Ulmer et al. (2022) turn such findings into a standard; we measure the distance between the standard and practice. The nearest ancestor in method is outside machine learning: Nuijten et al. (2016) recompute the $p$-values printed in psychology papers from the test statistics beside them, a corpus-scale re-test of published claims from the printed page. Marie et al. (2021) hand-annotate machine translation papers for evaluation practice and find the field's conventions drifting; they establish what careful manual annotation can see, at the scale manual annotation permits. Tang et al. (2020) show that the reporting practice proposed by Dodge et al. (2019) can itself mislead as usually implemented. We differ from all three in unit and scale: every highlighted comparison the pipeline recovers in 2895 papers, at the cost of reading only what is printed.

**Extracting results from papers.** Hou et al. (2019) and Kardas et al. (2020) extract result tables from machine learning PDFs to populate leaderboards: the same mechanics, the opposite purpose, since leaderboard extraction wants the winning number and can ignore a missing error bar. We know

of no released comparison-level annotation of error-bar presence against which an extractor could be scored; the paper-level confidence-interval label of Király et al. (2018) counts an interval anywhere in a paper, a different event, so Section 9 builds one.

## K THE PIPELINE, SPECIFIED

**Sampling.** Each year's corpus is the union of the NeurIPS and ICML accepted-paper lists as served by their open proceedings. A simple random sample without replacement is drawn with a fixed seed; PDFs are streamed, extracted and discarded, and each paper's SHA-256 is recorded so the corpus can be rebuilt and checked. Of the 2025 sample, 1059 papers are NeurIPS and 739 ICML. Their headline figures agree: 79.4% and 77.5% of comparisons print no error bar, 51.4% and 43.0% of testable ones fail, and 26.4% and 25.3% of papers print an explicit error bar somewhere.

**Text layer.** The extractor is PyMuPDF. Every text span is read with its font name, its font flags and its bounding box; the first 40 pages are read for tables and the whole document for the checklist and for run-count sentences. A span is bold when the font's bold flag is set or its name contains "bold". A font that is emboldened synthetically, or a Type 3 font whose flags are unset, is read as regular; Section 9 measures how often that happens by looking.

**Lines and tables.** Spans are clustered into lines by vertical overlap. A table is anchored on a span beginning "Table $N$": numeric cells within a fixed vertical window below the caption (and a small one above, for venues that caption underneath), within the caption's horizontal extent plus a margin, and forming a contiguous run of rows, are the table's cells. Cells past the next caption on the page belong to that caption. The horizontal constraint is what keeps a two-column layout's neighbouring table out; it was added after a sampled ICML table came back with the page number as its runner-up.

**Cells.** A cell is a number with an optional error bar in one of four forms: $a \pm b$, $a(b)$, $a(\pm b)$, or an interval $a\,[l, u]$. The parenthesised form is ambiguous and is credited as an error bar only when the caption or the surrounding page names a dispersion. What the bar means (a standard deviation, a standard error, a confidence interval or an interquartile range) is read from the caption and the page by keyword; a table whose bars are never named is "unlabelled" and is read as a standard deviation, the reading more favourable to the field.

**Run count.** A sentence of the form "$k$ (random) seeds/runs/trials" is sought in the caption first, then on the table's page, then anywhere in the paper; the caption wins if present, and the smallest count on the page, or the smallest anywhere in the paper, is taken when several appear, after a row label reading "top 100" was once read as a count. Table 2 gives the result under the caption-only reading and Section 8 the failure rate by source.

**Conversions and the test.** A bar labelled a standard deviation is divided by $\sqrt{n}$; a standard error is used as printed; a confidence interval's half-width is divided by 1.96; an interquartile range is not converted and its comparison is not testable. A rival that carries no bar makes the comparison untestable, whatever the winner carries. Welch's $t$ uses the Satterthwaite degrees of freedom $(s_A^2 + s_B^2)^2 / \big((s_A^4 + s_B^4)/(n-1)\big)$ with the stated $n$ on both arms, which at three runs is between two and four; one $n$ is read per table and applied to both arms. A standard error or interval with no run count is tested in the normal limit.

**Groups, orientation, direction.** Cells are grouped into columns by horizontal position and into rows by vertical position. The orientation is the grouping that yields more groups containing exactly one bold cell; a table yielding none is discarded. Within a group the bold cell is the winner and the best non-bold cell the runner-up; if the bold cell is the group's maximum, higher is better, if its minimum, lower, and otherwise the group is discarded as not a claim of victory. The test, Holm within table, and the bootstrap over papers are as stated in Section 2.

**What was excluded.** PDFs above 30 MB are skipped before download when the server states a size, and after it otherwise; 76 papers of the 2025 draw were excluded on that ground and 126 could

Table 9: Where tables and groups are lost between the PDF and the analysis, 2025 sample. Each discard is a decision that could have gone another way, so each is counted; the largest by far is tables in which no comparison group contains exactly one bold cell, most often because nothing is bold at all.

| | |
|---|---|
| Tables detected | 11106 |
| no comparison group with exactly one bold cell | 6530 |
| yielding at least one comparison | 4062 |
| Candidate comparison groups | 48004 |
| with exactly one bold cell and a rival | 79.0% |
| of those, bold cell not extremal (discarded) | 12.6% |
| extremal, tied with an unbolded cell (discarded) | 785 |
| extremal, in a table whose caption marks top three (discarded) | 53 |
| Highlighted comparisons analysed | 32300 |

not be fetched. Figure 2 bins by the improvement relative to the runner-up, in percent, with the last bin open; Section 7 reports gaps in points.