Can agent leaderboards tell neighbors apart? A paired reanalysis of the HAL rollouts

Taylor Kolasinski
Poisson Labs · taylor@poissonlabs.ai
October 22, 2026
DOI: [to be filled at publication]

Abstract

A leaderboard can be reliable as a whole and still unable to tell rank four from rank five. On eight of nine HAL benchmarks no neighboring pair of agents can be told apart on the tasks they share, and on most of them the gaps are smaller than anything the benchmark could detect. Read the board in ranges, not ranks.

Leaderboards are how most people choose an agent. A team picking a coding agent opens SWE-bench and takes whatever sits above the one they run now, and a lab launching a model leads with the place it landed. Both are asking the board the same question: is this agent better than that one?

The board looks like it answers that. It prints a score to the decimal with a rank beside it, and a higher rank reads as a better agent. We checked whether it can actually tell, and on these boards it mostly can't. In an exploratory redraw of one of them, 24 of 42 agents could claim first place. The public HAL rollouts behind a recent paper on leaderboard reliability contain 278 places where two agents sit next to each other on one of nine benchmarks. Compared on the tasks both of them ran, 3 of those 278 pairs show a difference a paired test can stand behind. All three are on one benchmark, USACO. On the other eight it is 0 of 265.

This is not a claim that the leaderboards are noise. The paper we built on, Hardy, Azam, Reuel, Kochenderfer and Koyejo's "Agent Evaluation Reliability: More Tasks Won't (Always) Fix An Agent Leaderboard", finds that rankings of whole systems (a model plus its scaffold) are highly reliable, with reliability coefficients between 0.935 and 0.994. We reproduced every HAL count in the paper from the authors' own released data, and nothing here disputes their numbers. A reliability coefficient asks whether the whole ordering would come out the same again. It does not ask whether the agent at rank four beats the agent at rank five. A board can answer the first question well and the second hardly at all, and that is what these boards do.

1 What this means if you read leaderboards

The coarse position on a board is information. The rank number mostly is not. Systems far apart usually do differ: on GAIA 373 of 561 pairs separate, and on CORE-Bench Hard 521 of 1,540 do. You just have to go a long way down. Among the top half of each board, the typical distance before a lower system is reliably worse is 7 places on GAIA, 23 on τ-bench Airline and 24 on CORE-Bench Hard. On USACO it is 2.

Most neighbor gaps are smaller than anything the benchmark could detect. An exact paired test cannot call a difference significant unless one system solves at least six more of the shared tasks than the other, and at exactly six it only clears if the other system wins none of them back. On a 50-task benchmark

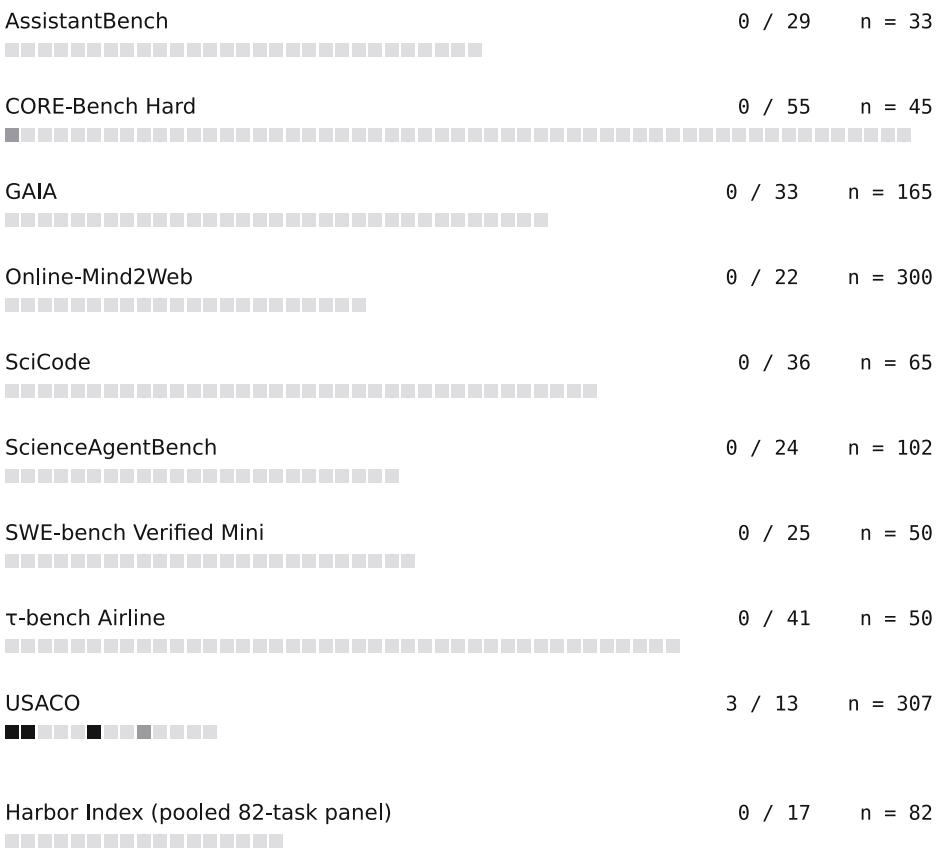


Figure 1. Adjacent-pair separability per benchmark. Each square is one pair of systems at neighboring printed ranks (r, r + 1), in rank order from the top of the board, so a row has $K - 1$ squares. Black: the pair separates under the primary rule, an exact two-sided McNemar test on the tasks both systems ran with Benjamini–Hochberg correction across the $K - 1$ adjacent pairs of that benchmark ($q < 0.05$). Grey: $p \leq 0.05$ uncorrected only. Light grey: neither. The count beside each benchmark is pairs separating under the primary rule out of its adjacent pairs, and n is its number of tasks. Across the nine HAL benchmarks 3 of 278 pairs separate, all on USACO (0 of 265 on the other eight; AssistantBench's pass/fail outcomes are the paper's thresholding of a partial-credit metric). Harbor Index (18 systems on 82 shared tasks) is the corroboration panel. Earliest run per replicated cell; with the latest run, 1 of 278.

that puts every gap under 12 points out of reach, however the tasks fall. 262 of the 278 neighbor gaps are under that floor. The floor is a limit, not a promise: a gap of six tasks can still fail the test.

Two of the nine benchmarks cannot rank anything. On AssistantBench and SciCode pass rates sit near 3%, and no pair of systems at any distance separates on either board, 0 of 435 pairs and 0 of 666. The largest neighbor gap on AssistantBench is 3 points against an 18-point floor. On SciCode it is 1.5 points against 9.

Task count and pass rates near the middle are what buy resolution here. USACO has 307 tasks and pass rates between 25% and 70%, and it is the only board where neighbors separate. That sits alongside the paper's title rather than against it. The paper's "more tasks won't fix it" is about ranking models across scaffolds, where task-by-scaffold interaction puts a ceiling on reliability. For telling two fixed systems apart, more tasks help, and most of these benchmarks have too few.

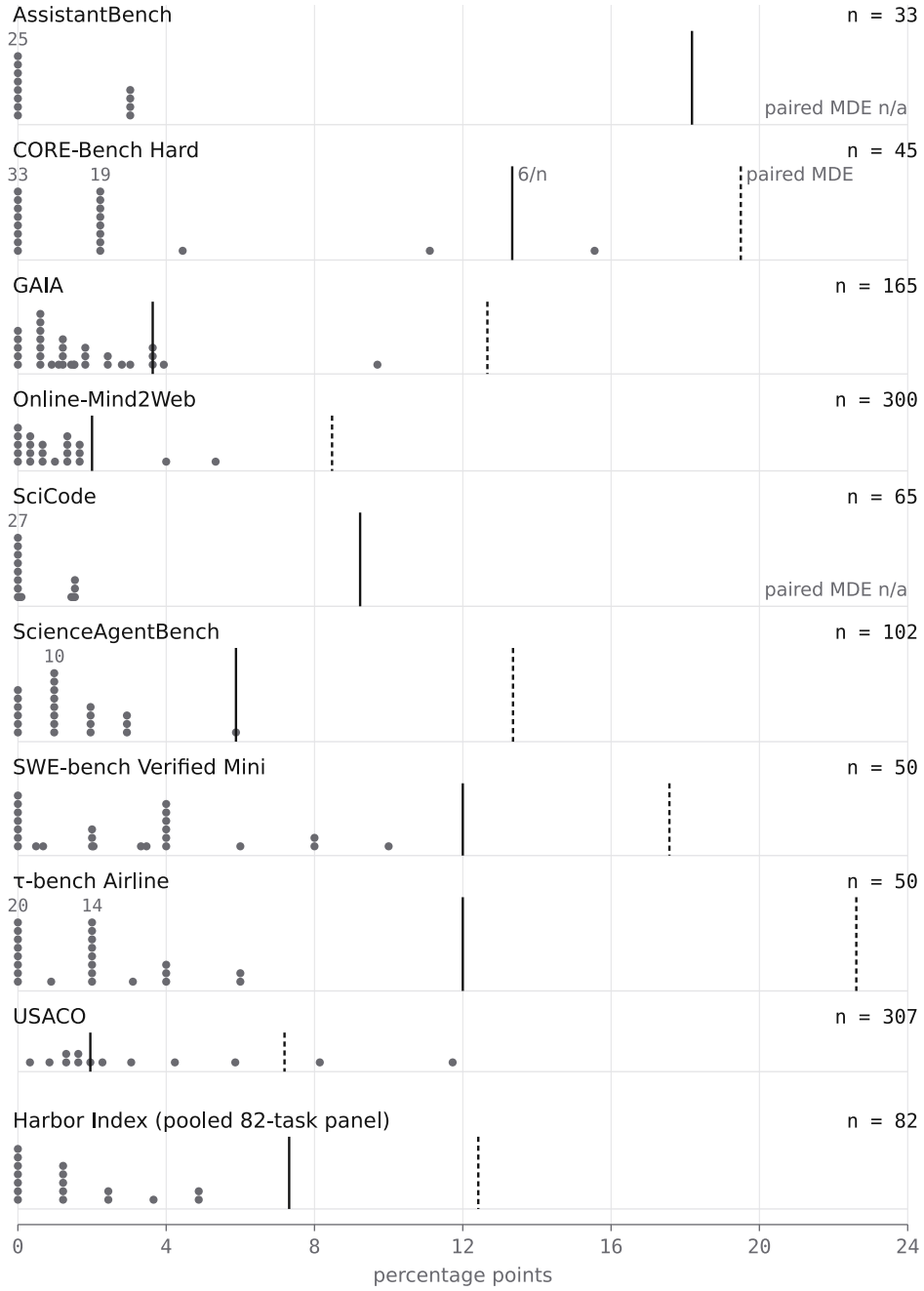


Figure 2. Resolution per benchmark. Each dot is the pass-rate gap between two systems at neighboring printed ranks, in percentage points; equal gaps stack, and a stack of more than eight is drawn eight high with its count above it. The solid line is the floor 6/n: an exact two-sided McNemar test at $\alpha = 0.05$ cannot reject unless one system solves at least six more of the shared tasks than the other, so no gap below 6/n can separate however the tasks fall. The dashed line is the paired detection limit (paired MDE): the smallest expected net difference at which that test has 80% power when each of the n tasks is discordant with probability ψ, the benchmark's median discordance between adjacent systems. On AssistantBench and SciCode no difference reaches 80% power at that ψ (n/a). n is the benchmark's number of tasks. 262 of the 278 HAL adjacent gaps lie below 6/n. Earliest run per replicated cell.

2 The same boards, redrawn

If a board printed only what its data supports, each agent would get a range of ranks instead of one number. The best rank an agent can claim is one plus the number of agents that clearly beat it. The worst is the size of the board minus the number of agents it clearly beats. This view is exploratory: it re-displays the same pre-registered tests and is not a new pre-registered result.

On SWE-bench Verified Mini, three agents can claim rank 1, all running sweagent, and none of the other 23 can. Below them, every agent's range overlaps the ranges of both its neighbors, and ten agents share the same worst rank of 26. The board can tell its top from its bottom and not much in between. (Handling the repeated runs the other way, with each system's latest run, six agents can claim rank 1.)

On other boards the top is more crowded. On τ-bench Airline 24 of 42 agents could be ranked first. On ScienceAgentBench it is 15 of 25, and on Online-Mind2Web 10 of 23. On AssistantBench and SciCode every agent's range is the whole board. USACO is the exception again: its top two agents are the only ones on any board pinned to a single rank.

That is what is broken about these boards, and it is a format problem more than a data problem. A board prints "1" for USACO's leader, whose place the data pins down, and prints the same "1" for τ-bench Airline's leader, which is one of 24 agents the data cannot order. Nothing on the page tells a reader which kind of "1" they are looking at. Every rank gets the same typeface and the same confidence, and the people choosing agents read them that way.

3 Rerun the same agent

Part of the reason is that one run is a noisy measurement. HAL ran 16 of its 287 systems more than once on the same tasks. Across those 18 pairs of runs, 14.5% of task outcomes flipped between runs (95% interval 12.7% to 16.5%). On the 6 run pairs whose ids differ only by a timestamp, which come from 4 systems, it was 18.1%. Four systems is a small base, so treat this as an order of magnitude and not a constant. At that rate, rerunning alone is worth roughly 7 points of standard error on the gap between two systems on a 50-task benchmark. A 3-point gap between two single runs is well inside what rerunning either agent could produce.

4 Scaffolds

The paper's sharpest warning is about scaffolds: the same model ranks differently depending on the harness around it. The task-level data shows the same thing from below. Of 122 swaps where one model was run under two scaffolds on the same tasks, 42 moved many tasks in both directions while the net score barely changed. After subtracting what rerunning alone does, 22 still do. That count is 25 if the three default-versus-high-effort run pairs and the one ambiguous pair are left out of the rerun baseline, and 13 if only the six timestamp-only reruns count.

Whether model A beats model B depending on the scaffold is real, and it is concentrated. It is confirmed by a direct test of the difference on three scaffold contrasts: GAIA's hal_generalist against hf_open_deep_research (46 or 47 of 78 model pairs, depending on the test), Online-Mind2Web's browser-use against seeact (24 of 45), and SWE-bench Verified Mini's hal_generalist against sweagent (1 of 1). On the other six benchmarks it is confirmed nowhere.

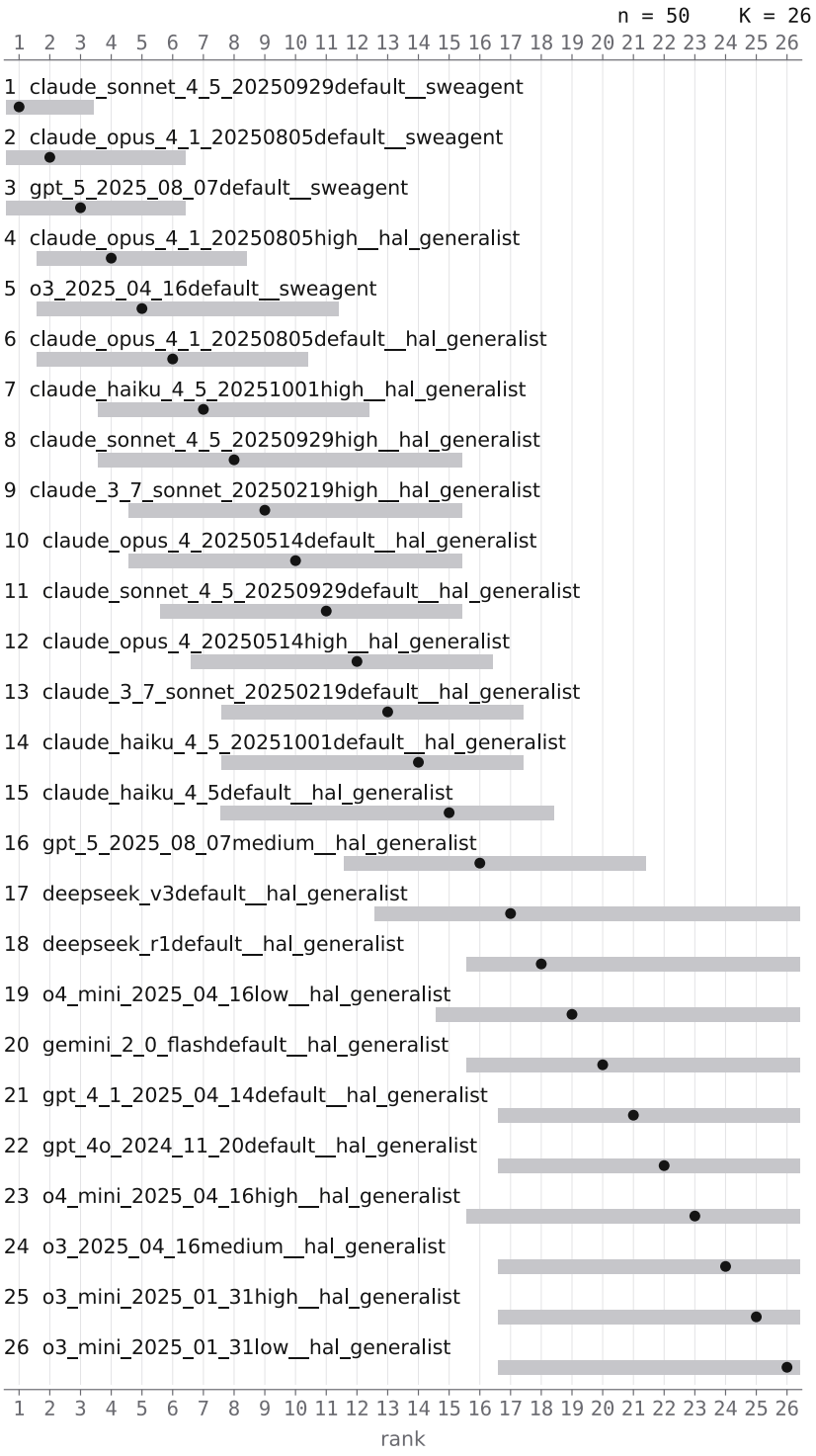


Figure 3. Rank ranges on SWE-bench Verified Mini (exploratory; not pre-registered). One row per system (K = 26, n = 50 tasks), in printed order: pass rate on the tasks the system ran, exact ties alphabetical. The dot is the printed rank. The bar runs from the system's best rank, one plus the number of systems clearly better than it, to its worst rank, K minus the number of systems it is clearly better than. Clearly better means the pair separates under an exact two-sided McNemar test on their shared tasks with Benjamini–Hochberg correction over all 325 pairs on the board ($q < 0.05$), and the system is the higher of the two on those tasks. A range is the set of positions the pairwise tests cannot separate, not a confidence interval on rank. The view re-displays the pre-registered all-pairs tests and is not a new result. Earliest run per replicated cell; with the latest run, six systems' ranges include rank 1 rather than three.



Figure 4. Scaffold-swap task churn against the rerun baseline. Each line is one within-model scaffold swap, the same model run under two scaffolds on the same tasks (122 on HAL, 9 on Harbor Index). The dot is the flip rate, the share of shared tasks solved under exactly one of the two scaffolds, and the line its 95% Wilson interval. The band is the same-system rerun flip rate with its 95% Wilson interval: the benchmark's own reruns where it has them, and hatched where it has none and the pooled rate over all 18 HAL run pairs from 16 systems stands in. Open circle: rules (a) and (b), no significant net change (exact McNemar, Benjamini–Hochberg across the benchmark's swaps, $q \geq 0.05$) with a flip-rate lower bound above 10%. Filled: also rule (c), the lower bound above the band's upper bound. Grey: neither. Counts at right: 42 HAL swaps meet (a)+(b) and 22 meet (a)+(b)+(c); with the rerun baseline leaving out the three default-versus-high-effort run pairs and the one ambiguous pair the second count is 25, and with timestamp-only reruns it is 13. Earliest run per replicated cell. Harbor Index has no reruns and uses the pooled HAL band.

5 What a leaderboard needs to answer that question

People bring one question to a board: does this agent beat that one. Answering it takes the outcome on every task, not only the score, and an honest statement of how small a gap the board can see. Most boards publish neither. Each fix below comes straight from a result above.

1. Publish per-task outcomes. The three pairs that separate here were found by comparing two agents on the same tasks, and a headline score cannot be compared that way.
2. Print the smallest gap each benchmark can detect next to its scores. 262 of the 278 neighbor gaps here sit below that floor, and a reader should see the floor before reading the rank.
3. Print a range of ranks for each agent instead of a single number. In the exploratory redraw above, that turns rank 1 on SWE-bench Verified Mini into a three-way claim, and on τ-bench Airline into a 24-way one.
4. Run the top entries more than once and publish every run. When 14.5% to 18.1% of task outcomes flip on a rerun, a gap between two single runs carries noise the board never shows.
5. Size a benchmark for the gaps it gets used to rank. USACO, at 307 tasks, is the only board here where neighbors separate.

None of this is expensive. It is a lot cheaper than a team choosing an agent because it sits one place higher.

6 What we ran

The data is the authors' own release, scaffold_eval at commit e7ec464: 29,923 HAL rollouts across 9 benchmarks, 1,117 tasks, 54 models and 13 scaffolds, making 287 systems. Every count in the paper's HAL tables came back identical. Harbor Index, the paper's second dataset, ran as a corroboration panel of 18 systems on 82 shared tasks, where 0 of 17 neighbor pairs separate.

Every comparison is paired on the tasks the two systems share, with an exact two-sided McNemar test, Benjamini-Hochberg correction across the neighbor pairs on each benchmark, and paired score intervals. Uncorrected, 5 of 278 neighbor pairs clear $p \leq 0.05$, and correcting across every pair on a benchmark instead of only the neighbors still gives 3. The tests ran through Poisson Prove, our paired-comparison tool. The few calculations Prove does not do yet, such as the paired detection limit and the scaffold interaction test, ran in the study code.

The analysis was pre-registered and frozen in a signed commit before any outcome was analyzed. Changes made after the data was reconciled but before any test ran are dated in the spec. Changes after the results are logged with their reasons, and none changes a verdict. About 4% of task cells hold more than one rollout. The primary analysis takes each system's earliest run, and repeating everything with the latest run gives 1 of 278 instead of 3.

A second model recomputed every headline number from the raw rows with its own code. Everything reproduced. The one count that depends on the choice of test, the GAIA scaffold interactions, is reported both ways above.

The whole read cost nothing to run. The study code rebuilds every table and figure from the released data with one command.

7 What this is a measurement of

Each comparison is on the shared tasks of two systems, one rollout per task and system, so it mixes which tasks were sampled with how a single run happened to go. It cannot separate the two. The rerun figure comes from 16 systems, and three of those 18 run pairs are default-effort and high-effort runs that the paper's matrix codes as one system. AssistantBench's metric is partial credit, and the pass/fail outcomes are the paper's thresholding of it. 39 of the 278 neighbor pairs are the same model at two effort levels or two scaffolds, distinct systems by the paper's definition, and none of them separates. The typical distance before systems separate is a description of the boards, not a tested quantity.

Cite this

```
@misc{kolasinski2026neighbors,
  author       = {Kolasinski, Taylor},
  title        = {Can Agent Leaderboards Tell Neighbors Apart? A Paired Reanalysis of the {HAL}
Rollouts},
  year         = {2026},
  month        = oct,
  organization = {Poisson Labs},
  howpublished = {\url{https://poissonlabs.ai}},
  note         = {Pre-registered paired reanalysis of the HAL rollouts released with Hardy et
al. (2026), arXiv:2610.00651}
}
```

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289–300.

Gelman, A., and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4), 328–331.

Hardy, M., Azam, R., Reuel, A., Kochenderfer, M., and Koyejo, S. (2026). Agent Evaluation Reliability: More Tasks Won't (Always) Fix An Agent Leaderboard. arXiv:2610.00651v1.

Hardy, M., et al. (2026). scaffold_eval: code and data released with arXiv:2610.00651. GitHub repository hardy-education/scaffold_eval, commit e7ec464 (MIT license). The HAL and Harbor Index matrices analyzed here.

Kapoor, S., Stroebl, B., Kirgis, P., et al. (2025). Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. arXiv:2510.11977v1.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

Shi, L., Lin, H., Zhu, Z., et al. (2026). Harbor Adapters and Harbor-Index: Infrastructure and a Curated Meta-Dataset for Large-Scale Agentic Evaluation. arXiv:2609.04298v3.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, 17(8), 891–908.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

Appendix: per-benchmark summary tables

benchmark	n_i	K	adjacent pairs	uncorrected p ≤ 0.05	BH adjacent (primary)	BH all pairs	Tango 95% excludes 0	all pairs separable (BH)	median d* top half	twin pairs (separable)	floor 6/n / max adjacent gap
AssistantBench	33	30	29	0	0	0	0	0 / 435	beyond table (15 of 15 censored)	10 (0 sep.)	18.2 pp / 3.0 pp
CORE-Bench Hard	45	56	55	1	0	0	1	521 / 1540	24.0 (0 of 28 censored)	4 (0 sep.)	13.3 pp / 15.6 pp
GAIA	165	34	33	0	0	0	0	373 / 561	7.0 (0 of 17 censored)	4 (0 sep.)	3.6 pp / 9.7 pp
Online-Mind2Web	300	23	22	0	0	0	0	119 / 253	8.0 (0 of 11 censored)	3 (0 sep.)	2.0 pp / 5.3 pp
SciCode	65	37	36	0	0	0	0	0 / 666	beyond table (18 of 18 censored)	12 (0 sep.)	9.2 pp / 1.5 pp
ScienceAgentBench	102	25	24	0	0	0	1	92 / 300	13.5 (0 of 12 censored)	2 (0 sep.)	5.9 pp / 5.9 pp
SWE-bench Verified Mini	50	26	25	0	0	0	0	211 / 325	5.0 (0 of 13 censored)	1 (0 sep.)	12.0 pp / 10.0 pp
tau-bench Airline	50	42	41	0	0	0	0	204 / 861	23.0 (0 of 21 censored)	3 (0 sep.)	12.0 pp / 6.0 pp
USACO	307	14	13	4	3	3	4	71 / 91	2.0 (0 of 7 censored)	0 (0 sep.)	2.0 pp / 11.7 pp
Harbor Index (pooled 82-task panel)	82	18	17	0	0	0	0	35 / 153	13.0 (4 of 9 censored)	4 (0 sep.)	7.3 pp / 4.9 pp

Table A1. Q1, adjacent-pair separability per benchmark. Primary analysis (earliest run per replicated cell); the last row is the Harbor Index pooled panel. n_i: tasks; K: systems; adjacent pairs: $K - 1$ systems at neighboring printed ranks. Every test is an exact two-sided McNemar test on the tasks both systems ran. Uncorrected: $p \leq 0.05$. BH adjacent (the primary count): Benjamini–Hochberg $q < 0.05$ over the $K - 1$ adjacent pairs. BH all pairs: adjacent pairs with $q < 0.05$ when the family is all $K(K-1)/2$ pairs. Tango 95% excludes 0: the Tango score interval on the paired difference excludes zero (descriptive; no verdict uses it). All pairs separable: pairs of any rank distance with $q < 0.05$ over all pairs. Median d* top half: for each rank r in the top half, the smallest d with (r, r + d) separable over all pairs, median over r (a rank with no such d counts as infinite; "beyond table" when the median is). Twin pairs: adjacent pairs of the same model at two efforts or two scaffolds. Floor 6/n: the smallest net difference an exact two-sided McNemar test at $\alpha = 0.05$ can ever call significant, as a share of n_i. Source: results/RESULTS.md.

benchmark	n_i	adjacent ψ min / median / max	paired MDE at median ψ	paired MDE over observed ψ range	paired MDE, n_d fixed at round(ψ n)	floor 6/n	unpaired MDE at p = 0.50	unpaired MDE at median system rate	unpaired MDE over observed rates	adjacent gap min / median / max	gaps < paired MDE	gaps < floor	gaps < unpaired MDE (0.50)
AssistantBench	33	0.000 / 0.061 / 0.152	n/a	n/a to n/a	n/a	18.2 pp	32.1 pp	24.3 pp (p = 0.030)	22.7 pp to 30.1 pp	0.0000 / 0.0000 / 0.0303	n/a / 29	29 / 29	29 / 29
CORE-Bench Hard	45	0.067 / 0.222 / 0.444	19.5 pp	n/a to 28.2 pp	18.5 pp	13.3 pp	28.0 pp	28.5 pp (p = 0.256)	19.3 pp to 29.1 pp	0.0000 / 0.0000 / 0.1556	55.0 / 55	54 / 55	55 / 55
GAIA	165	0.170 / 0.321 / 0.485	12.7 pp	9.3 pp to 15.6 pp	13.1 pp	3.6 pp	15.2 pp	15.3 pp (p = 0.474)	12.1 pp to 15.3 pp	0.0000 / 0.0121 / 0.0970	33.0 / 33	28 / 33	33 / 33
Online-Mind2Web	300	0.143 / 0.260 / 0.353	8.5 pp	6.3 pp to 9.9 pp	8.7 pp	2.0 pp	11.3 pp	11.1 pp (p = 0.320)	9.6 pp to 11.4 pp	0.0000 / 0.0067 / 0.0533	22.0 / 22	20 / 22	22 / 22
SciCode	65	0.000 / 0.031 / 0.094	n/a	n/a to n/a	n/a	9.2 pp	23.7 pp	15.0 pp (p = 0.031)	12.4 pp to 19.0 pp	0.0000 / 0.0000 / 0.0154	n/a / 36	36 / 36	36 / 36
ScienceAgentBench	102	0.078 / 0.221 / 0.304	13.4 pp	7.7 pp to 15.8 pp	13.7 pp	5.9 pp	19.1 pp	18.3 pp (p = 0.235)	8.8 pp to 19.3 pp	0.0000 / 0.0098 / 0.0588	24.0 / 24	23 / 24	24 / 24
SWE-bench Verified Mini	50	0.000 / 0.200 / 0.340	17.6 pp	n/a to 23.3 pp	16.7 pp	12.0 pp	26.7 pp	27.0 pp (p = 0.255)	15.7 pp to 27.7 pp	0.0000 / 0.0204 / 0.1000	25.0 / 25	25 / 25	25 / 25
tau-bench Airline	50	0.160 / 0.320 / 0.520	22.6 pp	15.7 pp to 29.3 pp	22.6 pp	12.0 pp	26.7 pp	27.4 pp (p = 0.440)	22.7 pp to 27.7 pp	0.0000 / 0.0090 / 0.0600	41.0 / 41	41 / 41	41 / 41
USACO	307	0.114 / 0.192 / 0.293	7.2 pp	5.6 pp to 8.9 pp	7.2 pp	2.0 pp	11.2 pp	11.2 pp (p = 0.387)	9.8 pp to 11.3 pp	0.0033 / 0.0195 / 0.1173	11.0 / 13	6 / 13	12 / 13
Harbor Index (pooled 82-task panel)	82	0.073 / 0.159 / 0.305	12.4 pp	n/a to 17.6 pp	12.0 pp	7.3 pp	21.2 pp	16.2 pp (p = 0.085)	13.2 pp to 21.2 pp	0.0000 / 0.0122 / 0.0488	17 / 17	17 / 17	17 / 17

Table A2. Q2, detection limits against observed adjacent gaps. Primary analysis; the last row is the Harbor Index pooled panel. ψ: the share of shared tasks on which two adjacent systems disagree. Paired MDE: the smallest expected net difference $\psi(2\pi - 1)$ at which the exact two-sided McNemar test at $\alpha = 0.05$ has 80% power when each of n_i tasks is discordant with probability ψ; n/a where no difference reaches 80% power. "n_d fixed" holds the discordant count at round(ψn). Unpaired MDE: the two-proportion normal-approximation MDE at 80% power, $\alpha = 0.05$, n_i per arm, at the stated pass rate. Gaps are pass-rate differences between adjacent systems (rates as shares); the last three columns count adjacent gaps strictly below each limit. Source: results/RESULTS.md.

board	n_i	K	separable pairs	median width	max width	single-rank ranges	ranges covering rank 1	printed rank outside own range
AssistantBench	33	30	0	29	29	0 of 30	30 of 30 (100.0%)	0
CORE-Bench Hard	45	56	521	37	51	0 of 56	2 of 56 (3.6%)	0
GAIA	165	34	373	12	14	0 of 34	3 of 34 (8.8%)	0
Online-Mind2Web	300	23	119	13	17	0 of 23	10 of 23 (43.5%)	0
SciCode	65	37	0	36	36	0 of 37	37 of 37 (100.0%)	0
ScienceAgentBench	102	25	92	19	21	0 of 25	15 of 25 (60.0%)	0
SWE-bench Verified Mini	50	26	211	9	13	0 of 26	3 of 26 (11.5%)	0
tau-bench Airline	50	42	204	33	38	0 of 42	24 of 42 (57.1%)	0
USACO	307	14	71	3	5	2 of 14	1 of 14 (7.1%)	0
Harbor Index (pooled 82-task panel)	82	18	35	13	17	0 of 18	6 of 18 (33.3%)	0

Table A3. Rank ranges per board (exploratory; not pre-registered). A re-display of the pre-registered all-pairs tests, not a new result. System X is clearly better than Y when the pair has Benjamini–Hochberg $q < 0.05$ over all $K(K-1)/2$ exact McNemar tests of the board and X is higher on their shared tasks. Best rank = 1 + the number of systems clearly better; worst rank = K − the number of systems it is clearly better than; width = worst − best. Earliest run per replicated cell. Source: results/RESULTS.md.