SAME RESULTS, DIFFERENT RANKINGS: MEASURING EVALUATION-VIEW DEPENDENCE IN LLM AGENT BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A benchmark can select different winners without changing any saved result. In DeepSWE, two configurations exchange first and twelfth place when the evaluation changes from single-attempt success to four-attempt task coverage. Such changes raise a measurement question: how much of a leaderboard depends on the evaluation view, and where does that dependence occur? We introduce the Structured Leaderboard Ambiguity Index (SLAI), which combines changes in model-pair relations with their ranking position and relative rank-gap amplitude across a declared family of views. Each contribution retains evidence from an actual pair of views. ParetoSweep computes the index exactly and returns these witnesses. Two theoretical controls establish information that selected conventional summaries omit: identical ranking summaries can accompany an SLAI ratio of $n - 2$, and duplicating an existing view leaves the entire ambiguity surface unchanged. Across ten saved evaluation panels, the full family can reveal substantially more weighted change than any fixed pair: BFCL yields 27.93% versus 9.36%. Four additional source panels retain both low-change cases and further family-coverage gaps. SLAI describes where view-dependent comparisons occur and preserves the evaluation conditions needed to inspect them.

1 INTRODUCTION

Frontier large language models (LLMs) are now released, compared, and marketed largely through their ranks on agentic benchmark leaderboards (Artificial Analysis, 2026a). Yet a model's rank can change without changing any outcome, even under metrics reported by the benchmark itself. Consider a recent example from DeepSWE (Huang et al., 2026), where each model attempts each of 113 tasks four times (Figure 1a). By single-attempt success rate (pass@1), GPT-6 Astra (xhigh) leads the leaderboard with 335 successes in 452 attempts, while GPT-5.5 (high), with 291, ranks twelfth. Under four-attempt success rate (pass@4), however, the two trade places: GPT-5.5 solves 102 tasks, Astra only 91. Such reversals are pervasive in current agentic evaluation: the same runs, interpreted through a different evaluation view, yield a different leaderboard (Figure 1b).

This reversal is not merely a presentation issue. A user who needs single-attempt reliability may choose Astra, whereas one who can retry may choose GPT-5.5. Claims such as "SOTA," "#1," or "beats X" likewise depend on which evaluation view is reported. Rankings also shape what developers optimize, so a poorly chosen view can steer research in the wrong direction. On AlpacaEval, for instance, models could climb the leaderboard simply by giving longer responses, until the benchmark adopted a length-controlled win rate as its default metric (Dubois et al., 2024). Similarly, after Arena accounted for answer length and Markdown usage, Claude 3.5 Sonnet rose to tie for #1 on hard prompts and Llama-3.1-405B climbed to #3, while GPT-4o-mini and Grok-2-mini fell below most frontier models. A ranking is therefore meaningful only when we know which evaluation view produced it.

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.

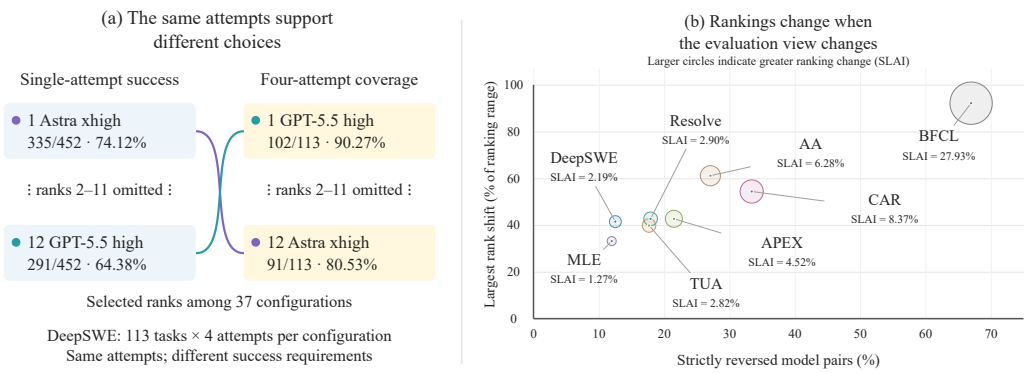

Figure 1: Same results, different leaderboards. (a) On DeepSWE, the first- and twelfth-ranked configurations swap places from pass@1 to pass@4. (b) Ranking changes across eight benchmarks between their officially reported views.

Prior work shows that rankings can shift with answer order, prompt weighting, scoring rules, or distributional assumptions (Alzahrani et al., 2024; Siska et al., 2024; Hua et al., 2025). Our core analysis starts from evaluation views already defined and reported by benchmark publishers, so ranking differences do not require a scoring rule introduced for this analysis; LiveBench's additional weighting is a separately identified analytical control. We ask: **how much model-comparison information does a single leaderboard omit across these views?** The Structured Leaderboard Ambiguity Index (SLAI) measures changes in pairwise relations, ranking positions, and rank gaps across the declared family, retaining the view pairs that support each contribution.

2 Related Work

Evaluation choices. LLM evaluation studies already document sensitivity to answer order, scoring and distributional assumptions. Alzahrani et al. (2024) report perturbation-induced ranking changes with normalized Kendall agreement. Siska et al. (2024) study prompt reweighting through performance distributions, rank changes and pairwise win frequencies. Hua et al. (2025) use mean Spearman correlation and performance dispersion to show that evaluation rules can account for apparent prompt sensitivity. We compare these reporting summaries on matched inputs (Section 5.4); our source-defined families also include retry and reliability requirements (Grace et al., 2026; Huang et al., 2026; Chen et al., 2026).

Ranking comparisons. Kendall measures pair-order agreement (Kendall, 1938). The broader ranking and information-retrieval literature also provides position-sensitive correlations and overlap measures (Vigna, 2015; Webber et al., 2010); Appendix E.5 gives a positive comparison. Our question concerns a benchmark's declared evaluation views: which relation changes, positions and amplitudes are jointly supported across the family? SLAI retains the responsible view pairs. Its separation results concern the stated conventional summaries, not every position-sensitive measure.

Families and evidence. Multiverse and specification-curve analyses retain defensible alternatives (Steegen et al., 2016; Simonsohn et al., 2020). Multiple-criteria ranking and partial identification distinguish necessary, possible and assumption-dependent conclusions (Greco et al., 2008; Tamer, 2010). Claim Replay connects evaluation assertions to evidence (Qin & Tong, 2026); evidence bounds and evaluation cards make support explicit (Gao & Zhou, 2026; Ghosh et al., 2026). We measure variation once view-specific answers are available. Witnesses retain their inputs' evidential depth: aggregate scores support comparisons at published precision.

3 Measuring Change Across Evaluation Views

Figure 2 follows the calculation from saved results to the comparisons that contribute to SLAI.

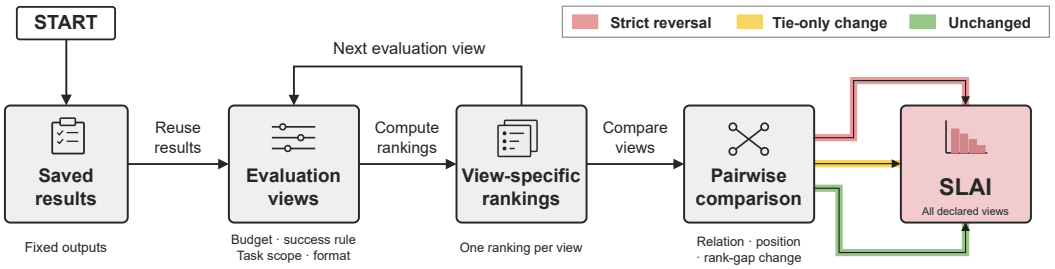

Figure 2: SLAI calculation. An actual view pair jointly supplies relation change, position and amplitude. Unchanged relations contribute zero but remain in the denominator. Saved results are reused; task or format views may originate from different outputs. No model is run by this calculation.

3.1 From saved results to a family of rankings

Let $n \geq 2$, $N = \binom{n}{2}$, and $\mathcal{Q}$ be a nonempty finite view family. Scores $s_i(q)$ are oriented so larger is better. Configuration identities stay fixed; task support may vary as declared, as in MLE's complexity groups. Missing cells require an explicitly restricted complete panel, not conversion to failures.

We use average ranks $r_i(q)$, with rank 1 best and tied configurations assigned the mean of the positions they occupy. Define the pair preference and normalized directed rank gap by

$$p_{ij}(q) = \begin{cases} 1 & s_i(q) > s_j(q), \\ 1/2 & s_i(q) = s_j(q), \\ 0 & s_i(q) < s_j(q), \end{cases} \qquad g_{ij}(q) = \frac{r_j(q) - r_i(q)}{n-1}. \tag{1}$$

The basic Leaderboard Ambiguity Index records the extent of relation change:

$$\mathrm{LAI} = \frac{1}{N} \sum_{i<j} \left[\max_{q \in \mathcal{Q}} p_{ij}(q) - \min_{q \in \mathcal{Q}} p_{ij}(q) \right]. \tag{2}$$

A pair contributes 1 if it strictly reverses somewhere in the family, $1/2$ if it changes only between a tie and one strict direction, and 0 if its relation stays fixed. For DeepSWE, $\mathrm{LAI} = (83 + 37/2)/666 = 15.24\%$. LAI is consequently different from the percentage of strictly reversed pairs, which is 12.46% here.

3.2 Position and amplitude need a common witness

For the same unordered view pair q, q', define

$$\begin{aligned} \delta_{ij}^{qq'} &= |p_{ij}(q) - p_{ij}(q')|, \\ h_{ij}^{qq'} &= 1 - \frac{\min\{r_i(q), r_j(q), r_i(q'), r_j(q')\} - 1}{n-1}, \\ a_{ij}^{qq'} &= \frac{|g_{ij}(q) - g_{ij}(q')|}{2}. \end{aligned} \tag{3}$$

Here δ measures relation change, h proximity to the top, and a change in directed rank gap. All lie in $[0, 1]$. Amplitude concerns ranks, allowing different score scales across views.

These choices encode a reporting preference. The minimum rank in h gives priority to a comparison if either configuration reaches a leading position in either view. Its linear scale avoids a discrete top-k cutoff. Directed gap change in a captures the extent of a reversal while remaining invariant to monotone score transformations. Uniform integration gives equal weight to equal threshold areas; it is a default, not an empirically optimal utility model. Family monotonicity, duplication invariance and witness retention hold for any fixed nonnegative threshold measure; numerical values and cross-panel order can depend on that choice (Appendix F).

With exactly two views, the construction is direct:

$$\mathrm{SLAI}(\{q, q'\}) = \frac{1}{N} \sum_{i<j} \delta_{ij}^{qq'} h_{ij}^{qq'} a_{ij}^{qq'} = d_{ha}(q, q'). \tag{4}$$

Unchanged relations contribute zero. A reversal contributes more near the top or with greater amplitude. A common witness is essential: multiplying separate maxima can assign a contribution that no comparison supports (Appendix D).

Figure 3 follows A, B across two six-configuration rankings. Their directed gaps are $1/5$ and $-3/5$, so $a = 0.4$; the relation reverses and reaches the top, giving $\delta = h = 1$. Their rectangle contributes 0.4 before averaging over all 15 model pairs.

3.3 A SURFACE OVER THE FULL FAMILY

For position threshold u and amplitude threshold v, define

$$\mathcal{A}(u, v) = \frac{1}{N} \sum_{i<j} \max_{q<q'} \left\{ \delta_{ij}^{qq'} \mathbf{1}[h_{ij}^{qq'} \geq u,\ a_{ij}^{qq'} \geq v] \right\}, \qquad \mathrm{SLAI} = \int_0^1 \int_0^1 \mathcal{A}(u, v)\, du\, dv. \tag{5}$$

The order $q < q'$ enumerates unordered view pairs; a maximum over no pairs is zero. Each model pair contributes the strongest qualifying change. Integrating weights position and amplitude; the denominator includes unchanged pairs.

The surface satisfies $\mathcal{A}(0, 0) = \mathrm{LAI}$ and $0 \leq \mathrm{SLAI} \leq \mathrm{LAI} \leq 1$, and is nonincreasing in either threshold. Uniform integration is an explicit preference for changes that qualify under stricter thresholds, not a probability of selecting the wrong model. Appendix F varies the weights.

Different model pairs or thresholds can have different actual witnesses. The envelope may exceed every fixed-pair distance, or the mean of each model pair's largest rectangle. Adding a view cannot decrease the surface at fixed configurations and weights; duplicating a weak ranking leaves it unchanged (Appendix A).

3.4 EXACT COMPUTATION WITH PARETOSWEEP

Each nonzero change defines a rectangle $[0, h] \times [0, a]$. For a model pair, let U_{ij}^{all} be the union of rectangles with $\delta \geq 1/2$, and U_{ij}^{strict} the union with $\delta = 1$. Its exact contribution is

$$e_{ij} = \tfrac{1}{2}\left(|U_{ij}^{\mathrm{all}}| + |U_{ij}^{\mathrm{strict}}|\right), \qquad \mathrm{SLAI} = \frac{1}{N} \sum_{i<j} e_{ij}. \tag{6}$$

ParetoSweep groups identical weak rankings, retains view aliases, enumerates view pairs, and scans widths in decreasing order while maintaining each layer's greatest height. Every increase retains its responsible witness; strip areas give the exact contribution.

Half-integer ranks permit rational arithmetic. For m views and m' distinct weak rankings, time is $O(mn \log n + N(m')^2 \log(m' + 1))$; the method still enumerates view pairs. Appendix D gives the scan and independent checks. The sweep uses standard rectangle-union operations; the contribution is their exact, witness-preserving application to this measure.

Connections to established distances. Ordinary preference distance $d = N^{-1} \sum \delta_{ij}$ satisfies $\overline{d} \leq d_{\max} \leq \mathrm{LAI}$. With two strict rankings, $\mathrm{LAI} = (1 - \tau)/2$; the amplitude-only moment is $M_{0,1} = \frac{n+1}{3(n-1)}(1-\rho)/2$. Configuration burdens $b_i = (n-1)^{-1} \sum_{j \neq i} e_{ij}$ satisfy $\mathrm{SLAI} = n^{-1} \sum_i b_i$. Appendix A proves these reductions, the zero criterion, invariances, decompositions and moment laws.

4 EXPERIMENTAL SETUP

The ten design and four additional panels contain 430 configuration/reference entries, 105 view columns, 2,383 cells and 12,966 within-panel pairs (Table 1). Models recur across sources, BBEH includes Random, and each panel has its own ranking universe.

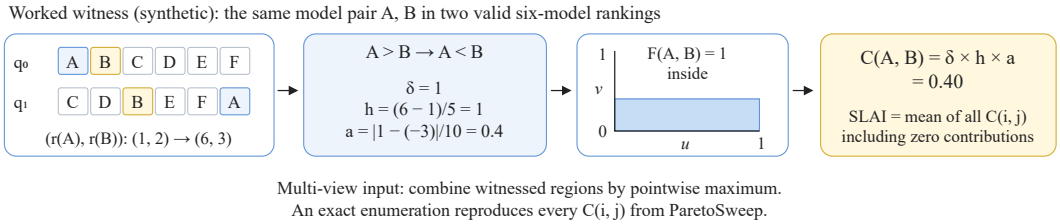

Figure 3: A worked witness in two six-configuration rankings. For A, B, the ranks change from $(1, 2)$ to $(6, 3)$, giving $\delta = 1$, $h = 1$ and $a = 0.4$. The shaded rectangle has area $e_{AB} = 0.4$. This is one pair's contribution; SLAI averages all 15 pair contributions, including zeros. With more views, rectangles combine by a pointwise maximum.

DeepSWE, TUA, Resolve, AA and APEX provide official success requirements; MLE/CAR vary scope and BFCL format. Banking is a two-configuration example; LiveBench adds analytical task weighting. EvalPlus, CRUXEval, LongBench and BBEH supply additional public data under locally locked defaults/candidate rules, with source eligibility clarified during acquisition (Appendix B).

We reuse published scores, trial labels or sufficient statistics at saved precision. Winners, stable relations, conventional summaries, LAI/SLAI and the strongest d_{ha} use matched inputs and weights. Location/duplication controls and a search over 383 view pairs test information differences. Sixteen weight and 25 tie settings assess sensitivity; independent grid integration checks ParetoSweep. No model, scorer or judge is rerun.

5 Results

5.1 Official views can change the choice and preserve many comparisons

DeepSWE's single-attempt and four-attempt questions select different unique winners in the complete 37-configuration panel (Figure 1a). Across its 666 model pairs, 83 strictly reverse and 37 change only through ties. The remaining 546 relations survive. The maximum within-configuration competition-rank span is 15 places out of 36; SLAI is 2.19%. A large change in the leading choice therefore coexists with substantial stable structure.

The pattern extends to other official choices. MLE's overall winner, Famou-Agent 2.0 / Gemini-3-Pro-Preview, gives way to AIBuildAI / Opus 4.6 on the High-complexity tasks. BFCL exhibits broader dependence: among 780 model pairs, 522 strictly reverse across format conditions, 23 change only through ties, and 235 remain fixed. Its LAI is 68.40%, while the strictly reversed fraction is 66.92%.

Stable choices also survive. Resolve retains one winner across single-attempt and all-eight success while preserving 22 of 28 relations; five reverse and one changes through a tie. LiveBench's supplementary weighting comparison preserves the winner and 1,572 of 1,596 relations, yielding SLAI 0.03270%. The added weighting thus produces a low-change control. Table 1 retains all ten panels, including the small Banking example, whose 100% describes one fully reversed pair.

5.2 Success concentration and partial credit explain different choices

Figure 4a explains the DeepSWE crossing. Astra's 44 extra successes concentrate on repeatedly solved tasks, while GPT-5.5 reaches 11 more tasks. Attempt success counts every success; coverage counts each reached task once. This saved-outcome arithmetic requires no hypothesis about internal reasoning.

TUA's GPT-5.5 xhigh has 314 full successes versus Opus 4.8 max's 323, yet partial credit reverses their mean-reward ordering (Figure 4b). The small mean difference describes saved outcomes without establishing a significant performance advantage.

Retry requirements differ again: GPT-5.5 covers 77 tasks and succeeds on all five attempts for 38; Opus covers 75 with 51 all-success tasks. TUA coverage has co-winners. Scope and format changes in MLE/BFCL cannot all be interpreted as rescoring identical generated outputs.

Table 1: Saved panels. LAI, SLAI and the strongest fixed-pair distance $d_{\max} = \max d_{ha}$ are percentages; ratios use unrounded design-panel values and saved five-decimal values for the additional sources. Icons identify code, tool-use, scope or task panels. *Analytical weighting control. †One model pair. BBEH includes Random; CRUXEval retains two publisher-marked unverified rows.

Saved panel			Relation	Position and amplitude		
Panel	n	m	LAI	SLAI	$d_{\max}$	Ratio
Ten design panels						
DeepSWE	37	2	15.24	2.19	2.19	1.00
TUA	21	4	20.00	2.82	2.64	1.07
Resolve	8	2	19.64	2.90	2.90	1.00
AA	32	3	28.83	6.28	6.15	1.02
MLE	28	4	14.29	1.27	0.83	1.52
APEX	8	3	21.43	4.52	4.52	1.00
CAR	12	9	37.88	8.37	4.61	1.81
BFCL	40	26	68.40	27.93	9.36	2.98
LiveBench*	57	2	1.50	0.03270	0.03270	1.00
Banking†	2	2	100.00	100.00	100.00	1.00
Four additional source panels						
EvalPlus	125	2	2.03	0.02963	0.02963	1.00
CRUXEval	43	2	10.52	0.68715	0.68715	1.00
LongBench	8	21	89.29	53.38921	31.14978	1.71
BBEH	9	23	91.67	56.78168	25.36350	2.24

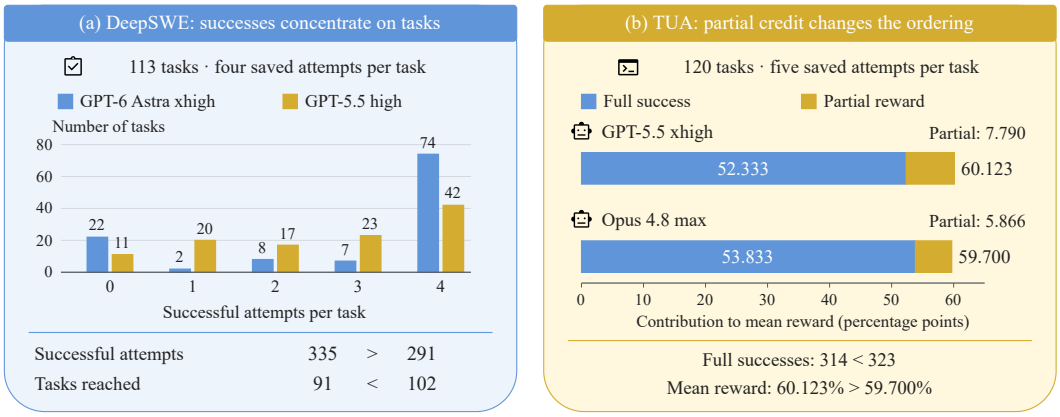

Figure 4: (a) DeepSWE task counts by successful attempts: each histogram sums to 113. Astra has more all-success tasks and more all-failure tasks. (b) TUA mean reward decomposes into full-success and partial-reward contributions (percentage points). GPT-5.5 xhigh has the larger mean despite fewer full successes than Opus 4.8 max. Exact values are in Appendix C.

5.3 The full family reveals changes that no fixed pair covers

For any family, the strongest fixed-view-pair distance is bounded by SLAI:

Proposition 1 (Fixed-pair lower bound)**.** *For fixed configurations and the weights in Equations 3–5,* $\max_{q<q'} d_{ha}(q, q') \leq$ SLAI. *With two views, equality holds.*

The family envelope contains each fixed pair's contribution at every threshold; integration gives the inequality. Figure 5 shows both equality and strict gaps.

Of BFCL's 325 view pairs, the strongest reaches 9.36486%, versus full-family 27.92649% (2.98 times). CAR/MLE yield factors 1.81/1.52. This gap has two sources. Different model pairs can require different view pairs as witnesses. Within one model pair, different witnesses can also cover complementary threshold regions.

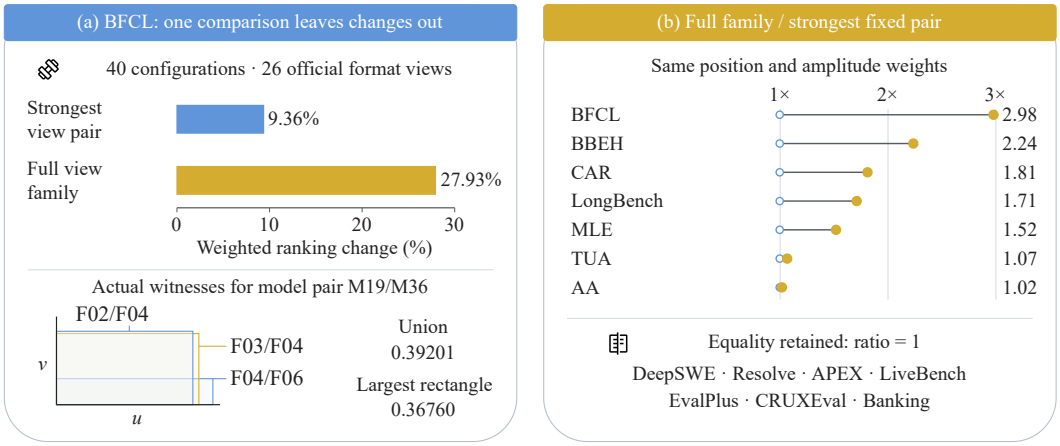

Figure 5: (a) BFCL's full-family SLAI exceeds its strongest fixed-pair d_{ha}. For the saved pair M19/M36, three rectangles have union $265/676$, exceeding the largest rectangle $497/1352$; these are pair contributions before averaging. (b) Ratios use identical position and amplitude weights. Equality cases remain visible; Table 1 reports their absolute values, including low-change controls. Cross-task families describe distinct task rankings.

The BFCL example in Figure 5a makes the second source concrete. Three comparisons strictly reverse M19/M36, but trade off proximity to the top against amplitude. Their rectangles cover a union of $265/676$, whereas the largest single rectangle covers $497/1352$. Retaining only that rectangle would discard witnessed change. These are contributions for one model pair, before averaging; the panel-level comparison additionally allows different model pairs to select different witnesses.

Equality distinguishes this coverage gain from the magnitude of change. Two-view families have only one candidate comparison, so equality is required. Their absolute values nevertheless range from EvalPlus's 0.02963% to Banking's 100%; Banking has only one model pair. APEX has three views and still retains equality: its extra view adds no area beyond the maximizing pair. More views therefore need not produce a strict family gap.

Additional LongBench/BBEH families yield factors 1.71/2.24; EvalPlus/CRUXEval retain low values and two-view equality. Cross-task rankings describe coverage across distinct tasks, not a controlled effect of rescoring one prediction set (Table 1).

5.4 Conventional summaries can agree while the location differs

Excluding Banking, SLAI has cross-panel Spearman correlation 0.700 with mean Kendall distance, 0.783 with mean Spearman distance and 0.950 with Kendall inconsistency. Association does not establish determination. Figure 6 and Table 2 test specific information differences; Appendix E retains the full identical-input comparison.

Theorem 1 (Location separation)**.** *For every $n \geq 4$, two pairs of strict rankings on the same configurations have identical Kendall and Spearman correlations, LAI, winner-set change, maximum rank displacement, and mean absolute rank displacement, while their SLAI ratio is $n - 2$. For $n \geq 7$, their Top-5 sets also remain unchanged.*

Swap positions 2/3 or $n - 1/n$ in the same order. Both reverse one pair with equal amplitude, but top proximity differs by $n - 2$. SLAI values are $(n - 2)/[N(n - 1)^2]$ and $1/[N(n - 1)^2]$. At $n = 100$, the ratio is 98 although both values and their difference approach zero. Appendix A gives the proof and approximation consequence.

CAR supplies a saved-data match: Base/single and Base/any-three, each compared with Hallucination/all-three, give identical ordinary distance, Kendall/Spearman correlations, tie-corrected Kendall inconsistency, maximum competition-rank span and winner Jaccard distance. SLAI is 2.20855% versus 1.97846%. This shared-endpoint match establishes non-redundancy for the six-summary vector, not its prevalence or a pure position effect.

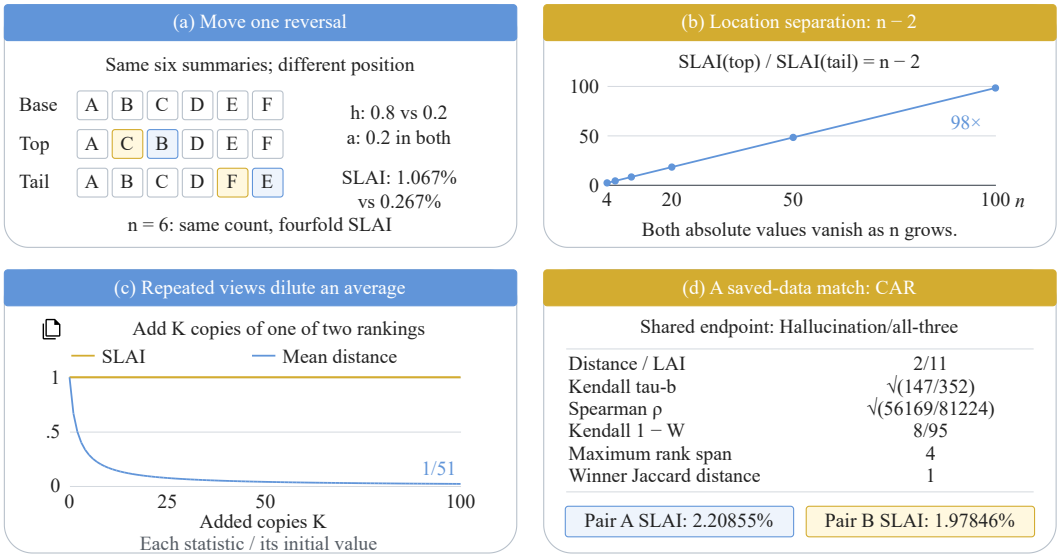

Figure 6: (a–b) Moving one adjacent reversal preserves Kendall, Spearman, LAI, winner change, maximum shift and mean absolute shift; the SLAI ratio is $n-2$. Both absolute values tend to zero as n grows. (c) An entry-weighted mean falls to $2/(K+2)$ of its initial value; SLAI stays fixed. (d) CAR's Base/single (A) and Base/any-three (B), each compared with Hallucination/all-three, match the six displayed summaries. Their SLAI values differ; this match does not isolate position alone (Appendix E).

Table 2: What the controlled comparisons distinguish. The first column concerns the strict-ranking construction in Theorem 1; the second concerns adding copies of one of two distinct rankings. A fixed-pair statistic requires an aggregation rule before duplication is meaningful.

Summary	Move reversal	Repeat a view
Kendall / Spearman	Unchanged	Depends on aggregation
Winner / max or mean shift	Unchanged	Depends on aggregation
Mean ordinary distance	Unchanged	Decreases as $2d/(K+2)$
Max ordinary distance	Unchanged	Unchanged
LAI	Unchanged	Unchanged
SLAI	Changes by factor $n-2$	Unchanged

The reporting summaries used in sensitivity studies also omit joint information. Two four-view constructions have identical rank distributions for every configuration, winner frequencies, pairwise win frequencies, and mean Kendall/Spearman correlations, yet SLAI is $2/3$ versus $7/9$ (Appendix E.4). Their configurations occupy different positions together across views. This comparison concerns the summaries, not the full retained ranking matrices or the validity of prior empirical findings. Position-sensitive rank measures supply further positive comparisons in Appendix E.5.

5.5 Repeated views add no ranking information

Theorem 2 (Duplication and averaging)**.** *Let two rankings have symmetric nonnegative discrepancy $d>0$, with identical rankings at discrepancy zero. After adding K copies of one ranking, the equally weighted mean over all unordered pairs of distinct entries is $2d/(K+2)$. The SLAI surface remains unchanged.*

There are $K+1$ cross-type pairs out of $\binom{K+2}{2}$. At $K=100$, the mean falls to $1/51$ of its original value. Added SLAI witnesses are repeated or have zero change. LAI, maximum distance and a deduplicated mean also avoid dilution. A probability-weighted mean remains appropriate when view frequencies carry substantive meaning.

5.6 THE MAIN STRUCTURE SURVIVES THE TESTED CALCULATION CHOICES

Table 3 summarizes the controls. BFCL, CAR, AA and APEX retain their leading order across tested weights, with LiveBench lowest; local changes involve Resolve, TUA and DeepSWE. BFCL's family gap survives all 25 tie settings. ParetoSweep and the independent grid agree exactly on 4,249 design-panel pairs. Appendices F and D give full policies and checks.

Table 3: Sensitivity and computation checks on saved panels. Weight correlations exclude Banking. The 20 random-priority settings are a subset of the 25 tie policies. Ranges describe sensitivity, not confidence intervals.

Check	Quantity	Observed result
Weights	Panel-order Spearman vs. default	$\geq .95$ (9 positive settings); $\geq .933$ (all 16)
Ties	BFCL family / fixed-pair ratio	2.92–3.05 (20 priorities)
Ties	DeepSWE SLAI	1.98–2.49% (25 policies)
Computation	Sweep vs. independent grid	Exact agreement on 4,249 pairs

6 IMPLICATIONS AND LIMITATIONS

Publishers can report which relations survive across declared views and which comparisons change. Developers can state whether a configuration leads on attempt success, coverage, repeatability or mean reward. Users can align that answer with their retry budget and requirement. LAI describes extent; SLAI emphasizes position and amplitude; witnesses locate the evidence. Low global SLAI can coexist with a consequential winner change, making shortlist-specific contributions useful.

Configuration membership, view scope and explicit weights affect the measure; SLAI is not a benchmark-quality or utility score. Format marginals cannot recover joint task success without aligned outcomes, and published ties may hide finer differences. These controls establish properties on saved inputs, not deployment utility, decision loss or population superiority.

7 CONCLUSION

A benchmark's official evaluation choices can support different model rankings on the same collection of saved results. Success concentration, partial credit, task scope, and format conditions explain why the supported comparison changes. SLAI measures the resulting relation changes through their position, amplitude, and coverage across the full view family, while retaining actual comparisons as witnesses. Theoretical controls and saved-panel analyses show information that selected conventional summaries and any single view pair can leave out. Reporting this structure alongside conditional winners preserves both what changes and what remains supported.

REPRODUCIBILITY STATEMENT

The anonymous supplementary package provides an offline path from saved public results to the reported rankings, SLAI values, witnesses, tables, and compiled manuscript. Its README gives dependencies and the single entry command `python reproduce.py -build-paper`; all 20 steps passed after fresh extraction. ParetoSweep agrees exactly with an independent integer-cell integration on all 4,249 model pairs in the ten design panels (Appendix D). Appendix G and Table 31 map each result to its saved inputs and scripts; Appendix A states assumptions and complete proofs, and Appendix F records the weight and tie controls. The package includes manuscript sources, editable figures, and an offline interactive DeepSWE demo (`demo/DeepSWE_Demo.html`); `python demo/rebuild.py` regenerates the demo from the packaged score table. Reproduction requires no new model calls or source downloads and concerns the saved results rather than rerunning the original evaluations.

AI USE STATEMENT

The original manuscript was written by the authors. Generative AI tools were used to polish the language and improve readability, not to write the manuscript in full. AI assistance was also used to prepare the accompanying source files, figures, and tables. The empirical analysis uses previously saved public evaluation results; preparing the manuscript did not involve additional model evaluations. The authors take responsibility for the final text, claims, and accompanying materials.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Artificial Analysis. Announcing the Artificial Analysis Intelligence Index v4.3, 2026a. URL https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-3. September 7, 2026.

Artificial Analysis. AA-AnalystAgent: Evaluation and methodology, 2026b. URL https://artificialanalysis.ai/evaluations/aa-analyst-agent. Public aggregate snapshot acquired September 12, 2026.

BBEH Authors. BBEH: Published task results, original paper table 2, 2026. URL https://arxiv.org/html/2502.19187v1. Saved September 19, 2026 from the original 2025 paper; eight model rows and Random across 23 tasks. Year denotes the source snapshot.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, 2024. doi: 10.18653/v1/2024.findings-emnlp.852. URL https://aclanthology.org/2024.findings-emnlp.852/.

Shoufa Chen, Luyuan Wang, Xuan Yang, Zhiheng Liu, Yuren Cong, Yuanfeng Ji, Feiyan Zhou, Xiaohui Zhang, Fanny Yang, and Belinda Zeng. TUA-Bench: A benchmark for general-purpose terminal-use agents, 2026. URL https://tuabench.ai/. Paper and official leaderboard; 21 Terminus-2 detail files acquired September 12, 2026.

CRUXEval Maintainers. CRUXEval: Official leaderboard and results CSV, 2026. URL https://crux-eval.github.io/leaderboard.html. Saved September 19, 2026; 43 published configuration rows. Year denotes the source snapshot.

Datacurve. DeepSWE v1.1: Public leaderboard, trial outcomes, and evaluation framework, 2026. URL https://deepswe.datacurve.ai/. Snapshot acquired September 12, 2026; v1.1 data generated September 3, 2026.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators, 2024. URL https://arxiv.org/abs/2404.04475. Version 2, March 10, 2025.

EvalPlus Maintainers. EvalPlus: Official results and metadata snapshot, 2026. URL https://github.com/evalplus/evalplus.github.io. Saved September 19, 2026; version 87418456d3a8c2bc6265f24ff91d2dd0e7993f4b. Year denotes the source snapshot.

Shanshan Gao and Liyi Zhou. Can agent benchmarks support their scores? evidence-supported bounds for interactive-agent evaluation, 2026. URL https://arxiv.org/abs/2605.10448v1. Version 1, May 11, 2026.

Avijit Ghosh, Anka Reuel, et al. Evaluation cards: An interpretive layer for AI evaluation reporting, 2026. URL https://arxiv.org/abs/2606.09809v1. Version 1, June 8, 2026.

Mikaela Grace, Jeremy Hadfield, Rodrigo Olivares, and Jiri De Jonghe. Demystifying evals for AI agents. Anthropic Engineering, 2026. URL https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents. January 9, 2026.

Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):416–436, 2008. doi: 10.1016/j.ejor.2007.08.013.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899, 2025. doi: 10.18653/v1/2025.emnlp-main.1006. URL https://aclanthology.org/2025.emnlp-main.1006/.

Wenqi Huang, Charley Lee, Leonard Tng, and Serena Ge. DeepSWE: Measuring frontier coding agents on original, long-horizon engineering tasks, 2026. URL https://arxiv.org/abs/2607.07946v1. Version 1, July 8, 2026; target definitions in Figure 1 and Section 5.3.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2), 1938. doi: 10.2307/2332226. URL https://doi.org/10.2307/2332226.

Johannes Kirmayr, Lukas Stappen, and Elisabeth André. CAR-bench: Evaluating the consistency and limit-awareness of LLM agents under real-world uncertainty. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40599–40618, 2026. URL https://aclanthology.org/2026.acl-long.1886/.

LiveBench Team. LiveBench: Public task scores and category mapping, release 2026-06-25, 2026. URL https://livebench.ai/. 57-configuration task-score snapshot acquired September 12, 2026.

LongBench Maintainers. LongBench: Original-v1 dataset descriptions and detailed results, 2026. URL https://github.com/THUDM/LongBench. Saved September 19, 2026; README at version 2e00731f8d0bff23dc4325161044d0ed8af94c1e. Year denotes the source snapshot.

Huanzhi Mao, Mingxuan Li, Ruoxi Wu, Shishir G. Patil, and Joseph E. Gonzalez. BFCL V4: Format sensitivity, 2025. URL https://gorilla.cs.berkeley.edu/blogs/17_bfcl_v4_prompt_variation.html. Official study released July 17, 2025, updated July 20, 2025; format-sensitivity data snapshot acquired September 12, 2026.

OpenAI. MLE-bench: Official leaderboard, complexity splits, and grading reports, 2026. URL https://github.com/openai/mle-bench. Saved main table and six grading reports acquired September 12, 2026; Famou-Agent 2.0 and AIBuildAI submissions.

Xi Qin and Jizhou Tong. What does an evaluation license? a commit-bound census of claim replay in Inspect Evals, 2026. URL https://arxiv.org/abs/2608.19269v4. Version 4, August 29, 2026; includes a portable reproduction package.

ResolveBench Team. ResolveBench: Public task-level success counts, 2026. URL https://resolvebench.com/. Saved data.js with 100 task IDs and eight configurations; snapshot acquired September 12, 2026.

Sierra. Official τ^3-Banking leaderboard and submission records, 2026. URL https://tau-bench.com/. Submission records dated August 4, 2026; snapshot acquired September 12, 2026.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560/.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Bertie Vidgen et al. APEX-Agents, 2026. URL https://arxiv.org/html/2601.14242v1. Version 1, Table 3; historical 480-task, eight-trial panel.

Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1166–1176, 2015. doi: 10.1145/2736277.2741088. URL https://doi.org/10.1145/2736277.2741088.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):1–38, 2010. doi: 10.1145/1852102.1852106. URL https://doi.org/10.1145/1852102.1852106.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, 2024. doi: 10.18653/v1/2024.findings-emnlp.108. URL https://aclanthology.org/2024.findings-emnlp.108/.

A PROPERTIES AND PROOFS

Throughout, the configuration set is fixed, scores are oriented consistently, ties use average ranks, and the view family is finite. A single-view family has no view pairs and hence zero LAI and SLAI. No sampling interpretation is needed for the following statements.

The appendix first develops the mathematical properties and their exact relationships to established distances, then defines the source views and works through the main cases. It next gives the integration algorithm, the complete comparison tables, the weight and tie analyses, and the evidence-removal controls. Tables contain saved results or explicitly labeled mathematical constructions.

A.1 THE INPUT OBJECT AND A COMPLETE NOTATION CONVENTION

The input is a rectangular score table together with a definition of each column. A row identifies a configuration, including the scaffold, model version, and run settings that distinguish it in the source. A column identifies an evaluation view and its task support. The same row identities must occur in every column of a panel. This requirement does not force different columns to use the same tasks: All and High in MLE explicitly use different supports. It instead prevents a missing configuration from silently changing the population being ranked.

Let $\mathcal{I} = \{1, \dots, n\}$ with $n \geq 2$, let $\mathcal{Q}$ be finite and nonempty, and let $N = \binom{n}{2}$. Scores $s_i(q)$ are oriented so that larger is better. At the saved source precision, define

$$b_i(q) = \sum_{k \neq i} \mathbf{1}[s_k(q) > s_i(q)], \qquad t_i(q) = \sum_{k} \mathbf{1}[s_k(q) = s_i(q)], \tag{7}$$

$$r_i(q) = b_i(q) + \frac{t_i(q) + 1}{2}, \qquad r_i^{\mathrm{c}}(q) = b_i(q) + 1. \tag{8}$$

Thus r_i is the average occupied position and r_i^{c} is competition rank. A tied group occupying positions 2, 3, and 4 receives average rank 3 and competition rank 2. SLAI uses r_i; reported rank spans use r_i^{c}. Keeping these conventions separate matters whenever a tied block changes size.

For each unordered configuration pair, choose an orientation $i < j$ and set

$$p_{ij}(q) = \begin{cases} 1 & s_i(q) > s_j(q), \\ 1/2 & s_i(q) = s_j(q), \\ 0 & s_i(q) < s_j(q), \end{cases} \qquad g_{ij}(q) = \frac{r_j(q) - r_i(q)}{n - 1}. \tag{9}$$

For one actual unordered view pair q, q', the complete witness is

$$\delta_{ij}^{qq'} = |p_{ij}(q) - p_{ij}(q')|, \tag{10}$$

$$h_{ij}^{qq'} = \frac{n - \min\{r_i(q), r_j(q), r_i(q'), r_j(q')\}}{n - 1}, \tag{11}$$

$$a_{ij}^{qq'} = \frac{|g_{ij}(q) - g_{ij}(q')|}{2}. \tag{12}$$

All three quantities belong to $[0, 1]$. Exchanging the two configurations replaces p by $1 - p$ and g by $-g$, leaving the witness unchanged. The orientation is consequently a bookkeeping device. A strict reversal has $\delta = 1$, a tie-to-strict change has $\delta = 1/2$, and an unchanged relation has $\delta = 0$. A change in the gap between two configurations that preserve their order is not, by itself, a contribution to this index.

Define the pair surface E_{ij} and its integral by

$$E_{ij}(u, v) = \max_{q < q'} \delta_{ij}^{qq'} \mathbf{1}[u \leq h_{ij}^{qq'},\ v \leq a_{ij}^{qq'}], \tag{13}$$

$$e_{ij} = \int_0^1 \int_0^1 E_{ij}(u, v)\, du\, dv, \qquad \mathcal{A}(u, v) = \frac{1}{N} \sum_{i<j} E_{ij}(u, v), \quad \mathrm{SLAI} = \frac{1}{N} \sum_{i<j} e_{ij}. \tag{14}$$

An empty maximum is zero. The view ordering in $q < q'$ only avoids duplicate comparisons. The finite family makes every surface a bounded step function, so these integrals exist and the interchange with the finite sum is exact. Boundaries have zero area under the default uniform measure; their inclusion remains explicit because the zero-exponent sensitivity variants put probability mass at zero.

What zero means. Under the default average-rank convention, $\mathrm{SLAI} = 0$ if and only if every configuration pair has the same weak relation under every view, equivalently $\mathrm{LAI} = 0$. To see the nontrivial direction, a changed relation gives a witness with $\delta > 0$. Its signed gap changes from positive to nonpositive, negative to nonnegative, or zero to nonzero, so $a > 0$. At least one endpoint is a strict comparison, whose better configuration has rank below n, so $h > 0$. Its rectangle has positive weighted area, forcing $e_{ij} > 0$. Zero therefore establishes agreement of the declared weak rankings. It says nothing about an omitted view, sampling uncertainty, or an unrecorded configuration.

All-tie columns are valid: every rank is $(n + 1)/2$ and every preference is $1/2$. Identical all-tie columns contribute zero. Comparing an all-tie column with a strict one can contribute through tie-to-strict changes. With one configuration, the normalization is undefined; such an input must be rejected or separately reported rather than silently assigned an SLAI. Missing scores likewise require an explicit scope decision before ranking.

A.2 Bounds, invariances, and family expansion

For a fixed model pair, p takes values in $\{0, 1/2, 1\}$. Its largest absolute difference over two views equals its maximum minus minimum over the family. At $(u, v) = (0, 0)$ all witness rectangles qualify, proving $\mathcal{A}(0, 0) = \mathrm{LAI}$. Increasing either threshold can only remove qualifying witnesses, so the surface is nonincreasing in each coordinate. Its range is contained in $[0, \mathrm{LAI}]$, which gives $0 \le \mathrm{SLAI} \le \mathrm{LAI} \le 1$ by integration over the unit square.

For any fixed view pair, the function $\delta \mathbf{1}[h \ge u, a \ge v]$ has integral $\delta h a$. The maximum defining each model pair's envelope is at least this function. Averaging and integrating proves Proposition 1. With two views there is only one candidate, giving equality. With additional views, equality can still hold when one fixed pair attains the full envelope contribution after integration.

Adding a view retains every existing candidate in the maximum, so it cannot decrease $\mathcal{A}$ or SLAI. This statement fixes the configurations and threshold measure: adding models changes ranks and normalization and is a different operation. If a new view has the same weak ranking as an existing view, its average-rank and preference columns coincide. Comparisons with other views repeat existing (δ, h, a) triples; comparing the copies gives $\delta = 0$. The entire surface is unchanged. This proves duplication invariance for any number of copies and any fixed integration measure.

Permuting configuration identities only permutes the model-pair summands. Reordering input rows with identities attached has the same effect. A strictly increasing transformation applied separately to a score column preserves its weak ranking, preferences, and average ranks. SLAI is invariant under these operations. The invariance concerns ranking information, so collapsed view aliases retain their distinct source meanings in the output.

A.3 Location separation

Theorem 1 uses a common strict ranking $A_1 \succ \cdots \succ A_n$. Let H exchange positions 2 and 3 and T exchange positions $n - 1$ and n. Each changes exactly one pair, moves two configurations by one place, and preserves the winner. Both have

$$\tau = 1 - \frac{4}{n(n-1)}, \qquad \rho = 1 - \frac{12}{n(n^2-1)}, \qquad \mathrm{LAI} = \frac{1}{N}. \tag{15}$$

The maximum displacement is 1, the mean absolute displacement is $2/n$, and the full multiset of absolute displacements is $\{1, 1, 0, \ldots, 0\}$. Winner-set Jaccard distance is zero. For $n \ge 7$, the H swap stays within the Top-5 and the T swap stays outside it, so both Top-5 sets remain unchanged.

The sole reversed pair in each case has $\delta = 1$ and $a = 1/(n - 1)$. Its highest position is 2 in H and $n - 1$ in T, giving

$$h_H = \frac{n-2}{n-1}, \quad h_T = \frac{1}{n-1}, \qquad \mathrm{SLAI}_H = \frac{n-2}{N(n-1)^2}, \quad \mathrm{SLAI}_T = \frac{1}{N(n-1)^2}. \tag{16}$$

Their ratio is $n - 2$, completing the proof. For $n = 4$ the same formulas apply; no Top-5 statement is needed. Table 4 gives a small instance that also controls the winner, unlike a comparison between a first–second swap and a last–second-last swap.

Table 4: Five-configuration position control. Both begin at $A \succ B \succ C \succ D \succ E$. The conventional summaries coincide while the weighted contribution differs.

	Exchange 2 and 3	Exchange 4 and 5
Second ranking	$A \succ C \succ B \succ D \succ E$	$A \succ B \succ C \succ E \succ D$
Kendall τ	$4/5$	$4/5$
Spearman ρ	$9/10$	$9/10$
LAI	$1/10$	$1/10$
Winner-set Jaccard distance	0	0
Maximum / mean absolute shift	1 / $2/5$	1 / $2/5$
SLAI	$3/160 = 1.875\%$	$1/160 = 0.625\%$

What the separation implies. No function of these six summaries and n can recover SLAI for every input, because the same summary vector has two distinct outputs. If one estimate x must approximate both positive values within factor $c \geq 1$, its feasible intervals intersect only if $\mathrm{SLAI}_H / c \leq c\,\mathrm{SLAI}_T$, hence $c^2 \geq n - 2$. There is no finite multiplicative guarantee independent of n from these summaries alone. At $n = 100$, the values are $98/48{,}514{,}950$ and $1/48{,}514{,}950$. Their ratio is large while their absolute values are small: the values scale as $\Theta(n^{-3})$ and $\Theta(n^{-4})$, and their absolute difference tends to zero. This is an information-separation result relative to the stated summaries, without a claim about empirical decision loss.

A.4 Duplication dilution

To prove Theorem 2, retain rankings q_A, q_B at discrepancy $d > 0$ and add K copies of q_A. The $K+1$ cross-type pairs each contribute d; all within-type pairs contribute zero. Under equal weighting of distinct input-entry pairs,

$$\overline{D}_K = \frac{(K+1)d}{\binom{K+2}{2}} = \frac{2d}{K+2}. \tag{17}$$

SLAI is unchanged by the weak-ranking duplication argument above. The same argument preserves LAI and the maximum pair distance. Deduplicating before averaging or holding each semantic view's total weight fixed avoids the dilution. Conversely, if repeated entries encode genuine probability mass, a frequency-weighted mean can appropriately change. The theorem distinguishes these aggregation questions.

A.5 When the full-family inequality is strict

Write $f_{ij}^{qq'}(u,v) = \delta_{ij}^{qq'} \mathbf{1}[u \leq h_{ij}^{qq'}, v \leq a_{ij}^{qq'}]$. For any fixed view pair $Q = (q, q')$,

$$\mathrm{SLAI} - d_{ha}(Q) = \frac{1}{N} \sum_{i<j} \int_{[0,1]^2} \big(E_{ij}(u,v) - f_{ij}^{Q}(u,v)\big)\, du\, dv. \tag{18}$$

Every integrand is nonnegative. Equality holds exactly when that same view pair agrees with every pair envelope almost everywhere. Since there are finitely many view pairs, $\max_Q d_{ha}(Q) = \mathrm{SLAI}$ exactly when at least one common view pair has this property. The condition explains both the mandatory two-view equality and the observed three-view equality in APEX. Merely having three or more views does not imply a positive family gap.

There are two places where taking a maximum can lose information. Different configuration pairs may need different endpoints, and a single configuration pair may need different endpoints in different threshold regions. These effects can be separated algebraically:

$$\max_Q \frac{1}{N} \sum_{i<j} \int f_{ij}^{Q} \;\leq\; \frac{1}{N} \sum_{i<j} \max_Q \int f_{ij}^{Q} \;\leq\; \frac{1}{N} \sum_{i<j} \int \max_Q f_{ij}^{Q}. \tag{19}$$

The first term is the strongest fixed-pair summary and the last is SLAI. The middle term lets each configuration pair choose its largest single rectangle, but still misses complementary rectangles within that pair. The BFCL calculation in Appendix D gives a concrete example of this second loss.

A.6 THE GENERAL ADJACENT-SWAP FORMULA

If a strict ranking is changed only by swapping positions $k, k+1$, exactly one configuration pair reverses. Its normalized signed gap moves from $1/(n-1)$ to $-1/(n-1)$, so $a = 1/(n-1)$, while $h = (n-k)/(n-1)$. Consequently

$$\mathrm{SLAI}(k) = \frac{n-k}{N(n-1)^2}, \qquad \mathrm{LAI}(k) = \frac{1}{N}. \tag{20}$$

There is one discordance among N pairs, which gives $\tau = 1 - 2/N$. The two nonzero rank displacements have squared sum 2, which gives $\rho = 1 - 12/[n(n^2-1)]$. Neither expression depends on k. This derivation identifies the source of the position separation without changing the number or size of rank displacements. Taking $k = 2$ and $k = n-1$ also preserves the winner. At $n = 4$, these are distinct swaps even though the two constructions involve a common configuration.

The non-identification claim is relative to the specified summary vector. Supplying the original rankings would, of course, restore the information needed to calculate SLAI. Adding another statistic that directly records the location of a reversal may also distinguish the examples. The theorem does not state that every possible rank statistic is insufficient.

A.7 DUPLICATION IN A GENERAL FAMILY

The two-view dilution theorem has a useful general form. Suppose an original family has m entries, pair-distance sum $S = \sum_{q<q'} D(q, q')$, and one selected view q_0 has distance sum $T = \sum_{q \neq q_0} D(q_0, q)$. If $D(q_0, q_0) = 0$, adding K exact copies of q_0 gives

$$\overline{D}_K = \frac{S + KT}{\binom{m+K}{2}}. \tag{21}$$

The numerator contains the original pairs and K new comparisons to each nonidentical entry. Pairs among the $K+1$ copies contribute zero. This mean tends to zero as K grows for any fixed finite family. For $m = 2$ it reduces to $2d/(K+2)$. If an observed family is a set of allowed interpretations, deduplication prevents this counting effect. If entries carry a declared sampling distribution, changing their mass is a different estimand.

A.8 DEPENDENCE ON THE CONFIGURATION POOL

SLAI is normalized within a panel. Adding configurations changes N, the ranks, and the denominators in h and a even if all old scores remain fixed. Thus the view-expansion monotonicity result cannot be applied to configuration expansion. The adjacent-swap formula already illustrates the distinction: extending a ranking by lower-ranked configurations changes the normalized value of the same local swap. A panel comparison should report n, the view family, and the support mapping alongside the index. A larger numerical SLAI is not an ordering of benchmark quality or an estimate of deployment regret.

A.9 DECOMPOSITION INTO CONFIGURATION BURDENS

Proposition 2 (Pair and configuration decompositions)**.** *Let* $e_{ji} = e_{ij}$ *and define* $b_i = (n-1)^{-1} \sum_{j \neq i} e_{ij}$. *Then*

$$\mathrm{SLAI} = \frac{1}{N} \sum_{i<j} e_{ij} = \frac{1}{n} \sum_{i=1}^{n} b_i. \tag{22}$$

Proof. Each unordered pair occurs twice in $\sum_i \sum_{j \neq i} e_{ij}$. Since $n(n-1) = 2N$, averaging b_i gives the pair average exactly. □

The quantity b_i records how much of the panel's relation variation involves configuration i, with the same location and amplitude conventions. It is neither an accuracy score nor a direction of superiority: a configuration can have high burden because it alternates between winning and losing. The contribution table can be used to select relations for inspection while keeping the original conditional scores available.

A.10 Ordinary distance, LAI, and Kendall correlation

Proposition 3 (The unweighted hierarchy and its two-view reduction)**.** *Let* $d(q, q') = N^{-1} \sum_{i<j} |p_{ij}(q) - p_{ij}(q')|$. *For at least two views,*

$$\overline{d} \leq d_{\max} \leq \mathrm{LAI}\,. \tag{23}$$

For two views, $\mathrm{LAI} = d(q, q')$. *If both rankings are strict, this value is* $(1 - \tau)/2$*, where* τ *is Kendall correlation.*

Proof. An average cannot exceed the maximum of its entries. For any fixed view pair, $|p_{ij}(q) - p_{ij}(q')| \leq \max_{v \in \mathcal{Q}} p_{ij}(v) - \min_{v \in \mathcal{Q}} p_{ij}(v)$. Summing proves the second inequality. With only two views, each pair's absolute difference equals its range. For strict rankings let D be the number of discordances. Then $d = D/N$ and $\tau = (N - 2D)/N$. □

The reduction is exact and establishes continuity with a familiar distance. With ties, ordinary d assigns half a unit to a strict-to-tie change, whereas Kendall τ_b also normalizes by the numbers of nontied comparisons in each ranking. The equality $d = (1 - \tau_b)/2$ therefore does not generally survive ties. For DeepSWE, $d = 15.24024\%$ while $(1 - \tau_b)/2 = 14.28344\%$ on the same two columns. They should not be interchanged.

A.11 The amplitude-only moment reduces to Spearman distance

Proposition 4 (Spearman reduction for two strict rankings)**.** *For two strict rankings of the same* n *configurations,*

$$M_{0,1} = \frac{\sum_i (r_i(q) - r_i(q'))^2}{n(n-1)^2} = \frac{n+1}{3(n-1)} \frac{1-\rho}{2}. \tag{24}$$

Proof. Relabel configurations so the first ranking has ranks $i = 1, \ldots, n$, and write π_i for their ranks in the second. Let $\mathcal{D} = \{(i, j) : i < j, \pi_i > \pi_j\}$ be its inversion set. Noninversions have $\delta = 0$. Every inversion has $\delta = 1$ and

$$a_{ij} = \frac{(j - i) + (\pi_i - \pi_j)}{2(n-1)}. \tag{25}$$

For a configuration i, let B_i count its smaller-π configurations to the right and A_i its larger-π configurations to the left. Counting configurations below π_i gives $B_i - A_i = \pi_i - i$. Therefore

$$\sum_{(i,j) \in \mathcal{D}} (j - i) = \sum_i i(A_i - B_i) = \sum_i i(i - \pi_i), \tag{26}$$

$$\sum_{(i,j) \in \mathcal{D}} (\pi_i - \pi_j) = \sum_i \pi_i(B_i - A_i) = \sum_i \pi_i(\pi_i - i). \tag{27}$$

Their sum is $\sum_i (i - \pi_i)^2$. Since removing the position threshold gives $M_{0,1} = N^{-1} \sum_{\mathcal{D}} a_{ij}$, division by $2N(n-1) = n(n-1)^2$ proves the first equality. Substituting the strict-ranking formula $\rho = 1 - 6 \sum_i (i - \pi_i)^2 / [n(n^2 - 1)]$ proves the second. □

This identity was checked in the saved analysis on all 120 permutations of five configurations. The proof applies to every $n \geq 2$ under its strict, two-view assumptions. Average-rank ties and the maximum over larger view families fall outside those assumptions. The amplitude component is consequently not presented as a new distance replacing Spearman; the full object adds location and jointly supported family coverage.

A.12 Moment reductions, order, and rectangle transformations

Proposition 5 (Threshold moments)**.** *For the measures defined in Appendix F,* $M_{0,0} = \mathrm{LAI}$*,* $M_{1,1} = \mathrm{SLAI}$*, and*

$$0 \leq M_{\alpha,\beta} \leq \mathrm{LAI}\,. \tag{28}$$

If $\alpha' \geq \alpha \geq 0$ and $\beta' \geq \beta \geq 0$, then $M_{\alpha',\beta'} \leq M_{\alpha,\beta}$. For two views and positive parameters,

$$M_{\alpha,\beta} = \frac{1}{N} \sum_{i<j} \delta_{ij} h_{ij}^{\alpha} a_{ij}^{\beta}. \tag{29}$$

For a general family, each layer is integrated by transforming its rectangle corners to (h^{α}, a^{β}) and taking the union area.

Proof. The point masses at zero evaluate $\mathcal{A}(0,0)$; the uniform measures give the defining SLAI integral. Bounds follow from $0 \leq \mathcal{A} \leq \text{LAI}$ and unit total measure. For $\alpha > 0$, a uniform variable U gives $U^{1/\alpha}$ with distribution function u^{α}; increasing α increases the threshold for every $0 < U < 1$. At $\alpha = 0$ use threshold zero. Couple the second coordinate in the same way. The surface is nonincreasing in each coordinate, proving the moment order. A rectangle has measure $h^{\alpha} a^{\beta}$. Monotone coordinate transformation preserves intersections and unions, so the layer-union calculation follows. With a single witness per configuration pair it reduces to the stated sum. □

In particular $M_{1,0}$ retains only location, and $M_{0,1}$ retains only amplitude. These are controlled ablations of the declared surface, rather than unrelated score functions. When a parameter is zero, the corresponding dimension is evaluated at zero; the implementation does not approximate it using a very small positive exponent.

A.13 A MAP OF THE THEORETICAL RESULTS

Table 5 collects the properties used in the paper, including all six properties in the saved SLAI definition. The distance reductions, exact integration, and matched constructions describe their relationship to the comparison measures. The more explicit equality condition and general-family duplication formula above unpack those arguments without introducing additional empirical claims.

Table 5: SLAI properties and the information each establishes. All invariances fix the configuration pool and threshold measure.

Property	Statement or condition
Bounds and zero	$0 \leq \text{SLAI} \leq \text{LAI} \leq 1$; zero iff all weak relations agree.
Threshold order	Raising either threshold cannot increase the surface.
View expansion	Adding views cannot decrease the surface or any fixed moment.
Duplicate invariance	Identical weak rankings preserve the entire surface.
Representation invariance	Relabeling and columnwise strictly increasing transforms preserve values.
Configuration decomposition	Pair contributions average to the same value as configuration burdens.
Fixed-pair bound	$\max d_{ha} \leq \text{SLAI}$; two views give equality.
Unweighted hierarchy	$\overline{d} \leq d_{\max} \leq \text{LAI}$; two strict views recover Kendall distance.
Spearman reduction	Two strict views give $M_{0,1} = \frac{n+1}{3(n-1)}(1-\rho)/2$.
Moment laws	Zero parameters remove thresholds; positive parameters transform rectangles.
Layer identity	Default integration is half the sum of two union areas.
Location separation	Six summaries can coincide while the SLAI ratio equals $n-2$.
Approximation obstruction	A common multiplicative approximation requires $c^2 \geq n-2$.
Duplication dilution	Equal-entry mean distance is $2d/(K+2)$ in the two-view construction.

B SOURCES, SCOPE, AND EVIDENCE DEPTH

B.1 TEN DESIGN PANELS

Table 6 identifies the declared families. A configuration is the model together with its scaffold and run settings, as represented in the saved source. It is not a model-name-only identity. Each panel keeps the same configuration rows across its views. A task subset can differ by view when the source definition explicitly specifies that difference.

The sources are the saved DeepSWE release (Datacurve, 2026), TUA detail files (Chen et al., 2026), Resolve success counts (ResolveBench Team, 2026), AA results (Artificial Analysis, 2026b), MLE

Table 6: Design-panel inputs. Counts are configuration rows by views. Evidence depth describes the ranking input, rather than a uniform level of trajectory access.

Panel	$n \times m$	Declared views	Saved evidence
DeepSWE	37×2	Single attempt; any of four	113 tasks, four binary labels per configuration and task
TUA	21×4	Mean reward; full single success; any of five; all five	120 tasks, five rewards, fixed Terminus-2 scaffold
Resolve	8×2	Single attempt; all eight	Per-task success counts for 100 tasks and eight attempts
AA	32×3	Single attempt; any of five; all five	Published chart reconstruction and exact fields
MLE	28×4	All 75; Low 22; Medium 38; High 15	Main-table aggregates; detailed reports for two systems
APEX	8×3	Single attempt; any of eight; all eight	Published aggregate table
CAR	12×9	Three task conditions, each with single, any-three, all-three	Published aggregate table
BFCL	40×26	26 format conditions	Per-format marginal scores
LiveBench	57×2	Equal category; equal task	Saved task scores; equal-task weighting is analytical
Banking	2×2	Single attempt; all four	Two saved official submissions

main table and grading reports (OpenAI, 2026), APEX and CAR tables (Vidgen et al., 2026; Kirmayr et al., 2026), BFCL format scores (Mao et al., 2025), LiveBench matrix (LiveBench Team, 2026), and Banking submissions (Sierra, 2026). The design-panel snapshots were acquired on 12 September 2026; source releases can be older.

DeepSWE's analysis uses the original 37-configuration complete panel. The broader 70-row archive contains incomplete attempt sets and is not pooled into this control. Single-attempt success averages all four saved trials, rather than selecting only trial one; any-four success requires recognition of a successful attempt. DeepSWE contributes $37 \times 113 \times 4 = 16{,}724$ labels and TUA contributes $21 \times 120 \times 5 = 12{,}600$ rewards. MLE's six detailed reports cover two configurations, 75 competitions, and three seeds, giving 450 records. These 29,774 directly processed records are not a count of independently reviewed trajectories.

Resolve's sufficient statistics summarize 6,400 attempts without supplying that many downloaded trajectories. AA's saved chart reconstruction is checked against 48 exact source fields, with no item-level verdicts. APEX and CAR are aggregate-table analyses. BFCL's format marginals do not bind a task-by-format correctness matrix and cannot determine all-format joint task success. Public-score rank calculations do not revalidate underlying model outputs, judges, or patch verifiers.

B.2 Four additional source panels

The SLAI definition, default $(\alpha, \beta) = (1, 1)$, comparison measures, and verification approach were fixed before the additional numerical analyses. The first local protocol was written on 19 September 2026 at 01:47:09.886 (UTC+8), before numerical EvalPlus retrieval began at 01:48:28.029. The second protocol was written at 02:09:14; its first numerical exposure was the LongBench README retrieval at 02:10:01.125. These timestamps establish the local order of operations on previously public data. The sources were not randomly sampled; previous incidental exposure cannot be excluded.

The four complete main scopes are:

- **EvalPlus HumanEval:** all 125 configuration rows in the bound official results version, with HumanEval and HumanEval+ scores (EvalPlus Maintainers, 2026).
- **CRUXEval:** all 43 rows in the selected publisher CSV, input- and output-prediction pass@1 (CRUXEval Maintainers, 2026). The two targets are distinct inference tasks. Two rows marked "under verification" by the publisher remain marked as such.

- **LongBench:** eight configurations across the 21 original-v1 detailed datasets (LongBench Maintainers, 2026). The family retains native F1, Rouge-L, accuracy, and edit-similarity scores, with task-specific supports. It does not represent a common success probability.
- **BBEH:** the eight language-model configurations and Random reference row in the original paper's Table 2, across 23 tasks (BBEH Authors, 2026). Complete here means complete for that published table. Additional overall-board configurations lack the same detailed task scores.

Some eligibility details could only be resolved after reading documentation or pages containing results. In particular, LongBench's native metrics, BBEH's table-specific scope, and L-Eval's comparability requirements are implementation decisions recorded in the admission materials. They are not represented as decisions completed before all numerical exposure.

B.3 Partial and auxiliary scopes

The candidate list and restricted results remain available. EvalPlus MBPP lacks 36 of the 125 configurations; its 89 complete cases supply auxiliary MBPP, average, and joint four- or six-view panels. BigCodeBench has 202 joined identities per condition, with only 153 complete Full/Hard pairs in Complete and 126 in Instruct. The missing 49 and 76 rows prevent either from entering the complete-source main set. L-Eval's topic-retrieval requirement for equal input length is not satisfied across the original family; a restricted 11-by-5 examination/aggregate scope is auxiliary. LongBench's 35-column and BBEH's 24-column extensions add published aggregates to their task columns and are not new sources.

Table 7: Retained auxiliary and partial scopes. These overlap the main data or fail the complete-family admission requirements and do not add independent sources. SLAI and $d_{\max}$ are percentages.

Scope	$n \times m$	SLAI	$d_{\max}$	Ratio
EvalPlus MBPP	89×2	0.11026	0.11026	1.000
EvalPlus average	89×2	0.06018	0.06018	1.000
EvalPlus joint four	89×4	0.95014	0.71951	1.321
EvalPlus joint six	89×6	0.95255	0.71951	1.324
LongBench 35 columns	8×35	53.64431	31.14978	1.722
BigCodeBench Complete	153×2	0.47647	0.47647	1.000
BigCodeBench Instruct	126×2	0.69250	0.69250	1.000
L-Eval restricted exam	11×5	7.41136	6.16591	1.202
BBEH tasks and overall	9×24	56.86849	25.36350	2.242

B.4 What each endpoint computes

For a task t with K saved attempts, write $y_{itk} \in \{0, 1\}$ for success and $c_{it} = \sum_{k=1}^{K} y_{itk}$. On a fixed support T, the three empirical endpoints are

$$s_i^{\text{single}} = \frac{1}{|T|K} \sum_{t \in T} c_{it}, \tag{30}$$

$$s_i^{\text{anyK}} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}[c_{it} > 0], \qquad s_i^{\text{allK}} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}[c_{it} = K]. \tag{31}$$

These are statistics of the saved attempts. In particular, any-K is observed task coverage, not the expression $1 - (1 - s_i^{\text{single}})^K$, which would add a homogeneous independent-trial assumption. Two configurations with the same mean success can have different coverage because their successes fall on different tasks. The single endpoint pools all trials in the balanced design; it does not select the first trial by file order.

DeepSWE uses single and any-four. Resolve uses single and all-eight. AA and APEX report single, any-K, and all-K for $K = 5$ and 8, respectively. CAR crosses three task conditions—Base, Hallucination, and Disambiguation—with single, any-three, and all-three. These formulas specify the

meanings of the reported endpoints; they do not imply that attempt records were obtained for every source. The evidence-depth table distinguishes direct reaggregation from use of published aggregates.

TUA additionally records rewards $z_{itk} \in [0, 1]$. Its mean-reward score is $\sum z_{itk}/(|T|K)$, whereas its binary endpoints use $y_{itk} = \mathbf{1}[z_{itk} = 1]$. Fractional rewards count in mean reward and do not count as complete successes. The publisher's "Success Rate" and "Pass@1" columns therefore answer different questions in this source. Their names alone would not suffice to implement the target mapping.

Task-support views. MLE's support sets are the 75 competitions and the disjoint Low, Medium, and High subsets of sizes 22, 38, and 15. The same 28 main-table configurations are retained across these columns. The detailed grading reports allow reaggregation for two systems and three seeds; they do not supply equally deep verification for all 28 rows. The All-versus-High comparison changes the support as specified by the source. It cannot be described as applying two formulas to an identical task subset.

LiveBench provides 23 task columns grouped into seven categories. If T_c is category c's task set and x_{it} is the saved task score, the views are

$$s_i^{\text{category}} = \frac{1}{7}\sum_{c=1}^{7}\frac{1}{|T_c|}\sum_{t\in T_c} x_{it}, \qquad s_i^{\text{task}} = \frac{1}{23}\sum_{t=1}^{23} x_{it}. \tag{32}$$

The first gives each category equal total mass; the second gives each task equal mass. Equal-task weighting is an analytical extension and is explicitly separated from publisher-defined alternatives. The model pool and saved task scores remain fixed in this comparison.

Format views. BFCL columns identify complete format conditions. Table 8 expands every F-ID used in the witnesses. The first 24 conditions cross return syntax, tool-call tag presence, and function-documentation syntax under plaintext/classic prompts. F25 and F26 are additional saved conditions. The analysis takes these columns as supplied, rather than inferring an unobserved full factorial design. Changing format can change the model's original responses; this is an analysis of saved conditional performance, not an intervention that only rescored identical predictions.

Table 8: Complete BFCL format dictionary from the saved source. Return and documentation columns give syntax; tag indicates whether a tool-call tag is present.

ID	Return	Tag	Documentation	Prompt	Style
F01	python	yes	python	plaintext	classic
F02	python	yes	xml	plaintext	classic
F03	python	yes	json	plaintext	classic
F04	python	no	python	plaintext	classic
F05	python	no	xml	plaintext	classic
F06	python	no	json	plaintext	classic
F07	json	yes	python	plaintext	classic
F08	json	yes	xml	plaintext	classic
F09	json	yes	json	plaintext	classic
F10	json	no	python	plaintext	classic
F11	json	no	xml	plaintext	classic
F12	json	no	json	plaintext	classic
F13	verbose_xml	yes	python	plaintext	classic
F14	verbose_xml	yes	xml	plaintext	classic
F15	verbose_xml	yes	json	plaintext	classic
F16	verbose_xml	no	python	plaintext	classic
F17	verbose_xml	no	xml	plaintext	classic
F18	verbose_xml	no	json	plaintext	classic
F19	concise_xml	yes	python	plaintext	classic
F20	concise_xml	yes	xml	plaintext	classic
F21	concise_xml	yes	json	plaintext	classic
F22	concise_xml	no	python	plaintext	classic
F23	concise_xml	no	xml	plaintext	classic
F24	concise_xml	no	json	plaintext	classic

ID	Return	Tag	Documentation	Prompt	Style
F25	python	no	json	markdown	classic
F26	python	no	json	plaintext	experimental

Configuration identity and decimal precision. Prompt-based and native function-calling variants remain distinct BFCL configurations. Effort levels and scaffolds remain distinct when represented as separate source rows. Published decimal scores are parsed as exact rationals at their displayed precision, so a displayed equality is a tie in this analysis. No hidden digits or uncertainty bands are reconstructed. Configuration identifiers, target labels, and source coordinates accompany the saved matrix, allowing a reader to distinguish genuine aliases from different experimental settings.

B.5 Interpreting the additional task families

The added panels broaden the types of declared views. EvalPlus contrasts HumanEval with its stronger HumanEval+ test suite on the same published configuration rows. CRUXEval contrasts input prediction with output prediction, retaining direct/chain-of-thought and budget distinctions in the source identities. LongBench retains all 21 task columns of the original v1 scope with each column's native score. BBEH retains the 23 task columns of the original paper's Table 2, including Random as a reference row. Within-column ranks make the calculation possible when columns have different score scales. They do not make F1, Rouge-L, edit similarity, and accuracy interchangeable units of utility.

Complete-case restriction can itself change a ranking problem. This is why MBPP's 89-row subset and the two BigCodeBench subsets are reported as auxiliary scopes rather than silently substituted for their original configuration families. Similarly, adding source aggregate columns to LongBench or BBEH expands an existing family; it does not add an independent benchmark observation. Low values, fixed-pair equalities, and unsuccessful admissions all remain in the saved results.

C How the Main Examples Produce Different Rankings

The examples below follow the data from task outcomes to target scores and then to pair relations. This separates the statistical reason for a ranking change from speculation about a model's internal behavior. DeepSWE and TUA have sufficiently detailed saved records to expose their endpoint decompositions; the other sources support the levels of analysis specified in Appendix B.

C.1 DeepSWE: repeated success and task coverage

For the opening comparison, Table 9 gives the number of tasks with exactly c successes among four attempts. Both rows sum to 113 tasks. The success total is the dot product with $(0, 1, 2, 3, 4)$, whereas coverage is 113 minus the first cell. For Astra, $2 + 2(8) + 3(7) + 4(74) = 335$ successes

Table 9: DeepSWE task counts by the number of successful attempts. Astra denotes mini-swe-agent / GPT-6 Astra xhigh; GPT-5.5 denotes mini-swe-agent / GPT-5.5 high.

Configuration	$c = 0$	1	2	3	4	Successes	Covered
Astra	22	2	8	7	74	335	91
GPT-5.5	11	20	17	23	42	291	102

and $113 - 22 = 91$ covered tasks. For GPT-5.5, $20 + 2(17) + 3(23) + 4(42) = 291$ successes and $113 - 11 = 102$ covered tasks. Thus

$$\Delta_{\text{single}} = \frac{335 - 291}{452} = \frac{11}{113} \approx 0.09735, \tag{33}$$

$$\Delta_{\text{any4}} = \frac{91 - 102}{113} = -\frac{11}{113} \approx -0.09735. \tag{34}$$

The differences are approximately $+9.735$ and -9.735 percentage points. The equal magnitudes are a feature of these saved counts, not an identity between the endpoints. Astra's successes are more

concentrated on tasks solved in all four trials: 74 such tasks compared with 42. GPT-5.5 covers more tasks with only one to three successes. This distribution accounts for the reversal without requiring an explanation in terms of hidden reasoning mechanisms.

Table 10: Task-level joint coverage of the two opening configurations; a task is covered when at least one of its four attempts succeeds.

	GPT-5.5 covered	Not covered	Total
Astra covered	88	3	91
Not covered	14	8	22
Total	102	11	113

The paired task table adds information absent from the two marginal histograms. The models both cover 88 tasks; Astra alone covers three and GPT-5.5 alone covers 14. Their coverage difference is therefore $3 - 14 = -11$ tasks. On eight tasks neither succeeds. This distinguishes broader coverage from a claim that one model's solved-task set contains the other's: the three Astra-only tasks rule out that containment.

The complete panel. The opening pair belongs to the original 37-configuration complete panel, containing 16,724 saved binary labels. Among its 666 configuration pairs, 83 strictly reverse and 37 change only through a tie; the remaining 546 preserve their relation. Hence

$$\mathrm{LAI} = \frac{83 + 37/2}{666} = \frac{203}{1332}, \qquad \mathrm{SLAI} = \frac{75667}{3452544}. \tag{35}$$

The first expression depends only on relation types; the second also uses their rank positions and gap changes. Table 11 lists all configurations and both competition ranks, so the opening reversal can be read in the context of the full panel. Configuration codes in this table shorten the source identifiers; they retain scaffold and effort distinctions. The largest competition-rank span is 15 places out of 36 possible, not the 11-place change of the opening pair.

Table 11: All 37 DeepSWE configurations. Scores are percentages; ranks are competition ranks. The source identifier components are retained with spaces for readability; mini denotes mini-swe-agent.

Configuration	Single	Rank	Any 4	Rank
mini / gpt 6 astra xhigh	74.12	1	80.53	12
mini / gpt 6 astra high	73.23	2	82.30	8
mini / gpt 6 astra max	73.23	2	79.65	17
mini / gpt 6 astra medium	72.79	4	82.30	8
mini / gemini 3 8 flash medium	71.02	5	83.19	6
mini / claude fable 5 xhigh	69.91	6	88.50	2
mini / grok 4 6 medium	67.48	7	84.07	5
mini / gpt 5 5 xhigh	67.04	8	88.50	2
mini / gpt 6 astra low	67.04	8	79.65	17
mini / gemini 3 7 flash medium	65.49	10	83.19	6
mini / gemini 3 7 flash high	65.27	11	82.30	8
mini / gpt 5 5 high	64.38	12	90.27	1
mini / deepseek v4 pro max	62.83	13	88.50	2
mini / gpt 5 6 sol medium	61.06	14	80.53	12
mini / gpt 5 6 terra xhigh	60.18	15	80.53	12
mini / gpt 5 6 luna xhigh	56.86	16	78.76	21
mini / muse spark 1 2 xhigh	54.87	17	81.42	11
mini / gpt 5 5 medium	53.98	18	77.88	22
mini / gemini 3 7 flash low	53.76	19	77.88	22
mini / gpt 5 6 terra high	53.76	19	80.53	12
mini / grok 4 5 high	53.76	19	77.88	22
mini / deepseek v4 flash max	53.32	22	80.53	12
mini / muse spark 1 1 xhigh	53.32	22	79.65	17
mini / claude opus 4 8 high	51.77	24	77.88	22
mini / gpt 5 4 xhigh	51.77	24	77.88	22

Configuration	Single	Rank	Any 4	Rank
mini / claude opus 4 8 medium	48.67	26	76.11	27
mini / claude sonnet 5 high	48.23	27	79.65	17
mini / gemini 3 6 flash high	46.68	28	75.22	28
mini / gpt 5 6 sol low	45.35	29	71.68	30
mini / gpt 5 6 luna high	44.25	30	75.22	28
mini / glm 5 2 high	36.28	31	68.14	31
mini / gemini 3 5 flash high	36.06	32	63.72	32
mini / kimi k2 7 code default	30.53	33	61.06	33
mini / gpt 5 5 low	26.99	34	47.79	34
mini / gemini 3 1 pro preview high	11.73	35	28.32	35
mini / gpt 5 6 luna medium	11.28	36	27.43	36
mini / gpt 5 6 luna low	1.55	37	4.42	37

C.2 TUA: MEAN REWARD, COMPLETE SUCCESS, AND RELIABILITY

TUA fixes 120 tasks and five attempts per configuration. The distinction between fractional progress and complete success is visible in the leading configurations. For any reward $z \in [0, 1]$,

$$z = \mathbf{1}[z = 1] + z\mathbf{1}[0 < z < 1]. \tag{36}$$

Averaging yields an exact decomposition of mean reward into the complete-success contribution and the partial-credit contribution. Table 12 applies it to GPT-5.5 xhigh and Opus 4.8 max, and retains the Opus 4.7 max coverage co-winner. GPT-5.5 exceeds Opus 4.8 by 0.4235 percentage points in mean

Table 12: TUA reward decomposition and repeat-attempt outcomes. The three reward columns are percentage points; the last two columns count tasks out of 120. All use Terminus-2.

Configuration	Mean	Full	Partial	Any 5	All 5
GPT-5.5 xhigh	60.12300	52.33333	7.78967	77	38
Opus 4.8 max	59.69950	53.83333	5.86617	75	51
Opus 4.7 max	57.98233	51.00000	6.98233	77	47

reward. Its complete-success contribution is 1.5 points lower, while its partial-credit contribution is 1.9235 points higher. The net difference follows directly from these two components. No significance claim is attached to the small mean-reward difference.

The repeat-attempt endpoints reveal a second distinction. GPT-5.5 completely solves at least one attempt on 77 tasks and all five attempts on 38. Opus 4.8 covers 75 tasks and solves all five on 51. Opus 4.7 also covers 77 tasks, making coverage a shared maximum in the 21-row panel. The stored histograms, ordered by the number of complete successes from zero through five, are

$$\text{GPT-5.5: } (43, 4, 5, 10, 20, 38), \tag{37}$$

$$\text{Opus 4.8: } (45, 2, 8, 6, 8, 51), \tag{38}$$

$$\text{Opus 4.7: } (43, 11, 4, 8, 7, 47). \tag{39}$$

These counts identify why coverage and all-five reliability select different configurations. They do not identify a causal effect of model family or effort because the compared systems are not a randomized intervention.

C.3 STABLE WINNERS CAN COEXIST WITH CHANGING RELATIONS

LiveBench's equal-category and equal-task views retain the same champion, while 24 of its 1,596 configuration pairs reverse and 1,572 remain stable. There are no tie changes in this panel. Its $\mathrm{LAI} = 24/1596 = 2/133$ and $\mathrm{SLAI} = 1091/3336704$ describe those changed relations even though winner-set change is zero. The smaller weighted value indicates that the changed pairs occupy a small total position–amplitude area. A stable winner and a stable complete ranking are distinct empirical findings.

Resolve similarly retains its leading configuration across single-attempt and all-eight success while changing other relations. Its input is the per-task success count, which is sufficient for these endpoints:

summing the counts gives single-attempt success, and counting tasks with eight successes gives all-eight success. The order of successful attempts and their execution trajectories are unnecessary for those two calculations, but would be needed for different questions about temporal dependence or failure modes.

MLE illustrates a support change. All 75 and High 15 compare the same 28 configuration identities on different explicitly declared competition subsets. A change in the winner supports a conditional performance statement about those subsets. It does not establish that changing an aggregation formula alone caused the difference. This support distinction is carried into the target registry, alongside the narrower two-system scope of the detailed grading reports.

D Exact Computation and Independent Checks

D.1 Integer coordinates and the two-layer identity

Set $R_i(q) = 2r_i(q)$, an integer even with average-rank ties. Define

$$\begin{aligned} D_h &= 2(n-1), & D_a &= 4(n-1), \\ H &= 2n - \min\{R_i(q), R_j(q), R_i(q'), R_j(q')\}, \\ A &= \left|(R_j(q) - R_i(q)) - (R_j(q') - R_i(q'))\right|. \end{aligned} \tag{40}$$

Then $h = H/D_h$ and $a = A/D_a$. At any threshold location, the maximum relation grade is 0, $1/2$, or 1. It equals one half of the indicator that any nonzero-change rectangle covers the location plus one half of the indicator that a strict-reversal rectangle does. Integrating proves Equation 6.

D.2 ParetoSweep

The following scan is applied separately to each model pair after collapsing identical average-rank columns and retaining their view aliases.

1. Enumerate the distinct view pairs. Retain $(H, A, 2\delta, q, q')$ when $\delta > 0$.
2. Group records by H and visit distinct widths in decreasing order. Append a final width of zero.
3. Maintain B_{all}, the largest A seen for $\delta \geq 1/2$, and B_{strict}, the largest A seen for $\delta = 1$.
4. At each width, update both heights and retain the responsible view pair whenever a height increases. Accumulate $(H - H_{\text{next}})(B_{\text{all}} + B_{\text{strict}})$.
5. Divide the accumulated value by $2D_hD_a$ to obtain e_{ij}. Average over all N model pairs, including zero-contribution pairs.

Within each width strip, the union heights are constant and equal to the maintained maxima. The strips do not overlap and cover every rectangle, proving exactness. Maintaining separate layers prevents a tie-change rectangle from obscuring a strict reversal. Working space, excluding output, is $O(mn + (m')^2)$ when model pairs are processed sequentially; retained contributions and witnesses require additional output space.

For BFCL, the saved run enumerates 253,500 model-pair/view-pair candidates, contains 60,974 nonzero rectangles, and retains 1,844 layered frontier points. These are computational objects, not independent observations. A single view pair may appear in both layers. Duplicating the first view 50 times raises the unreduced candidate count to 2,223,000; collapsing identical weak rankings leaves the enumerated candidate count at 253,500. No comparative runtime claim is made from this reduction.

D.3 Independent enumeration

An independent verifier fills the finite $D_h \times D_a$ grid directly from all real view pairs. Each cell receives the maximum covered grade $2\delta \in \{0, 1, 2\}$. Its contribution is

$$e_{ij}^{\text{grid}} = \frac{\sum_{x=1}^{D_h} \sum_{y=1}^{D_a} c_{xy}}{2D_hD_a}. \tag{41}$$

Every rectangle boundary lies on this integer grid, so this is exact integration rather than a sampling approximation. It does not use a Pareto frontier, width sorting, or the scan's area decomposition. Rational equality is checked for each model pair before comparing panel totals.

The saved checks agree on all 4,249 model pairs in the ten design panels. Further saved analyses check 129,639 model-pair contributions across 255 design-panel/tie-policy and EvalPlus scope combinations, 20,589 across eight second-batch main or auxiliary scopes, and 132 in the CAR matched comparisons. These sum with the initial 4,249 to 154,609 comparison operations, with repeated use of configurations and data. They are not an empirical sample size. Parameter variants with quarter-unit tie penalties use a general integer-grade integration, rather than applying the default two-layer formula outside its assumptions.

Controls also cover all-tie and single-view inputs, model relabeling, row permutation, strictly increasing score transforms, duplicate rankings, small-grid rectangle unions, and the mathematical constructions. The unbounded separation theorem follows from its proof, not finite numerical enumeration.

D.4 Why the two-layer scan preserves the maximum

The default relation grades are $0, 1/2, 1$. At a point covered by no nonzero witness, both layer indicators are zero. At a point covered by tie changes but no strict reversal, only the all-change layer is one and the contribution is $1/2$. At a point covered by a strict reversal, both indicators are one and the contribution is 1. These three exhaustive cases prove the layer identity pointwise, including points where multiple rectangles overlap. Adding rectangle areas directly would double-count those overlaps.

Within a layer, a rectangle with corner (H, A) can be removed if another has width at least H and height at least A. It is entirely contained in the other rectangle. This dominance operation must respect the layers: a wide, tall tie-change rectangle does not authorize deleting a smaller strict-reversal rectangle from the strict layer. Equal-width records must be processed as a group before accumulating the next strip. Retaining any one source witness for each attained height is sufficient for area calculation; aliases and alternative witnesses can also be kept for interpretation.

A small tied example. Take three configurations with rankings $A \succ B \succ C$ and $A \sim B \succ C$. Only the A, B relation changes. Its ranks are $(1, 2)$ and $(1.5, 1.5)$, giving $\delta = 1/2$, $h = 1$, and $a = 1/4$. Thus $e_{AB} = 1/8$ and $\mathrm{SLAI} = (1/8)/3 = 1/24$. LAI is $(1/2)/3 = 1/6$. The all-change layer has area $1/4$ and the strict layer is empty, producing the same $1/8$ pair contribution. Counting this event as a full reversal would double it; ignoring ties would erase it.

D.5 Executable specification of the scan

The following pseudocode makes the update order explicit. Each record carries the configuration IDs, source view IDs, and score-cell coordinates in the implementation output.

```
rank every score column using average occupied positions
collapse identical rank vectors; retain all source aliases
total := 0
for each unordered configuration pair (i,j):
  records := all nonzero (H,A,grade,view-pair) witnesses
  widths := distinct H values, decreasing, followed by 0
  height[all] := 0; height[strict] := 0; area := 0
  for adjacent widths (w,next):
    process every record with H = w
    update all height; update strict height if grade = 2
    retain the source witness when a height increases
    area += (w-next) * (height[all]+height[strict])
  contribution := area / (2 * D-h * D-a)
  save contribution, including zeros; total += contribution
return total / number-of-configuration-pairs
```

The scan invariant is that, after processing width w, each maintained height is the largest height among all rectangles in that layer with width at least w. In the open strip between the next width and

w, these are exactly the rectangles that cover the horizontal coordinate. The invariant is initially true with empty layers and remains true by a maximum update. Multiplying strip width by its constant union height is therefore exact. The disjoint strips exhaust the union up to boundaries of area zero, proving correctness of the algorithm.

Ranking costs $O(mn \log n)$ with sorting. If m' distinct weak rankings remain, each configuration pair has at most $\binom{m'}{2}$ candidate witnesses. Sorting and scanning them yields a total worst-case time

$$O\big(mn \log n + N(m')^2 \log(m'+1)\big)\,. \tag{42}$$

The implementation enumerates view pairs; the frontier is a compression of the evidence and an integration device. This complexity statement does not claim a faster asymptotic algorithm than every alternative rectangle-union method. Source aliases are excluded from the geometry but retained in the evidence output.

D.6 A COMPLETE BFCL UNION CALCULATION

Table 13 uses the same M19/M36 pair as the main example. Their saved average-rank gaps are 30 at F02, 29 at F03, $-5/2$ at F04, and 9 at F06. The three listed view pairs all strictly reverse that relation, but trade off position and amplitude. Under F03, the two configurations score 82.5 and 35.5 with average ranks 4.5 and 33.5; under F04 they score 79.5 and 80.0 with ranks 15.5 and 13. These source cells determine the F03/F04 witness directly. The format dictionary shows that the two views differ in tool-call tag presence and function-documentation syntax. Scanning the three widths gives

Table 13: Three actual BFCL witnesses for Qwen3-30B-A3B-Instruct-2507 (Prompt) and ToolACE-2-8B (FC). Coordinates use the full 40-configuration panel.

View pair	δ	h	a	ha
F04 / F06	1	1	$23/156$	$23/156$
F03 / F04	1	$71/78$	$21/52$	$497/1352$
F02 / F04	1	$34/39$	$5/12$	$85/234$

$$e_{19,36} = \left(1 - \frac{71}{78}\right)\frac{23}{156} + \left(\frac{71}{78} - \frac{34}{39}\right)\frac{21}{52} + \frac{34}{39}\frac{5}{12} \tag{43}$$

$$= \frac{265}{676} \approx 0.392012. \tag{44}$$

The saved full-family frontier for this pair yields this same value. Its largest single rectangle is $497/1352 \approx 0.367604$, while multiplying independently selected maxima gives $1 \times 5/12 \approx 0.416667$. The first loses supported complementary regions; the second introduces an unsupported upper-right region. The joint-witness union lies strictly between them.

At thresholds $(u, v) = (0.90, 0.40)$, F03/F04 qualifies because $71/78 > 0.90$ and $21/52 > 0.40$. F04/F06 fails the amplitude requirement; F02/F04 fails the position requirement. A single maximizing witness can therefore change with the threshold. The pair contribution above is not the full-panel SLAI: it is one of 780 terms whose average is $5301029/18982080$.

D.7 GENERAL RELATION GRADES AND INDEPENDENT VERIFICATION

For grades in $\{0, 1/L, \ldots, 1\}$, let U_ℓ be the union of rectangles whose grade is at least ℓ/L. The general layer identity is

$$\max_Q \delta_Q \mathbf{1}_{R_Q} = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}_{U_\ell}. \tag{45}$$

At a point whose maximum grade is k/L, exactly k layer indicators are one. Integrating proves the identity. The default uses $L = 2$; the quarter-unit tie penalties use $L = 4$. The independent grid checker instead writes the largest integer grade directly into each covered cell. This avoids borrowing the sweep's frontier or its strip decomposition.

Table 14: Relation-type partition over the full view family, recomputed from the saved score table. Strict counts pairs with both preference directions; tie-only excludes such pairs.

Panel	Pairs	Strict	Tie-only	Stable
DeepSWE	666	83	37	546
LiveBench	1596	24	0	1572
TUA	210	37	10	163
Resolve	28	5	1	22
AA	496	134	18	344
MLE	378	45	18	315
APEX	28	6	0	22
CAR	66	22	6	38
BFCL	780	522	23	235
Banking	1	1	0	0

For a pair with a strict reversal somewhere in the family, its LAI width is one even if other views tie. A tie-only pair has width one half. Thus the relation counts supply an independent accounting identity, $\mathrm{LAI} = (n_{\mathrm{strict}} + n_{\mathrm{tie}}/2)/N$, separate from the area integration. A stable pair contributes zero to both indices even if its rank gap changes across views. Keeping all zero-contribution pairs in N prevents the result from becoming an average only among affected relations.

E Conventional Summaries and the CAR Match

E.1 Precisely which conventional summaries are compared

All comparison measures use the same configuration rows and declared score columns within a panel. Their outputs describe different aspects of these rankings. Reporting their values together does not create a common accuracy objective on which a larger value is better.

Kendall and Spearman. For a view pair, let C, D count strict concordances and discordances, and let $T_q, T_{q'}$ count pairs tied only in the indicated view. Pairs tied in both are omitted from these four counts. We compute

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_q)(C + D + T_{q'})}}. \tag{46}$$

Spearman ρ is Pearson correlation of the two average-rank vectors. We report the means of $(1-\tau_b)/2$ and $(1-\rho)/2$ over all unordered view pairs. A zero denominator is undefined, not zero discrepancy. All view-pair correlations in the ten saved design panels are defined; all-tie inputs in a general implementation still require this rule. The exact Kendall and Spearman reductions in Propositions 3 and 4 explain when parts of the SLAI family coincide with these established quantities.

Kendall concordance. Let $R_i = \sum_q r_i(q)$, $\bar{R} = m(n+1)/2$, and $T_q = \sum_G (|G|^3 - |G|)$ over tied groups G in view q. The tie-corrected coefficient is

$$W_K = \frac{12 \sum_i (R_i - \bar{R})^2}{m^2(n^3 - n) - m \sum_q T_q}. \tag{47}$$

We report $1 - W_K$. Here T_q denotes the usual tie correction, distinct from the one-sided tie count in the preceding paragraph. A zero denominator is again undefined. W_K is a multi-ranking concordance statistic and should not be conflated with the count or identity of winning configurations.

Winner and Top-5 summaries. The champion set is $W(q) = \{i : r_i^{\mathrm{c}}(q) = 1\}$. The descriptive winner-change column is the mean of $\mathbf{1}[W(q) \neq W(q')]$. The matched CAR search instead uses winner-set Jaccard distance $1 - |W(q) \cap W(q')|/|W(q) \cup W(q')|$. These are distinct conventions when winner sets overlap without being identical. The Top-5 set includes the entire tied block at the fifth occupied position. Its comparison is the mean Jaccard distance over view pairs; if $n < 5$, the whole configuration pool is included. Thus Banking has zero Top-5 distance although its two configurations reverse completely.

Rank displacement. The maximum family span is $\max_i[\max_q r_i^c(q) - \min_q r_i^c(q)]/(n-1)$. For the strict two-view synthetic controls, maximum and mean absolute displacements are computed directly between their two ranks. These summaries retain extremes or aggregate movements, without locating each reversed relation. The theorem holds the full absolute-displacement multiset fixed, which is stronger than matching its maximum alone.

Table 15: Identical-input comparison of relation and correlation summaries on every design panel. Values are percentages of the corresponding normalized quantity.

Panel	LAI	SLAI	Mean d	Max d	Kendall	Spearman
DeepSWE	15.24024	2.19163	15.24024	15.24024	14.28344	7.17963
LiveBench	1.50376	0.03270	1.50376	1.50376	1.50376	0.16204
TUA	20.00000	2.81897	10.35714	19.04762	9.66365	4.25804
Resolve	19.64286	2.89723	19.64286	19.64286	19.08580	10.47833
AA	28.83065	6.27919	19.22043	28.12500	18.83687	11.82159
MLE	14.28571	1.26970	7.49559	10.44974	6.83309	2.10350
APEX	21.42857	4.51895	14.28571	21.42857	13.80512	7.35855
CAR	37.87879	8.36777	14.35185	25.00000	13.17209	6.48729
BFCL	68.39744	27.92649	22.12465	36.73077	21.63215	14.51980
Banking	100.00000	100.00000	100.00000	100.00000	100.00000	100.00000

Table 16: Concordance, leading-set changes, and weighted fixed-pair comparisons on the same inputs. Winner change is an indicator averaged over view pairs; Top-5 uses Jaccard distance. All values are percentages.

Panel	$1 - W_K$	Winner	Top-5	Mean d_{ha}	Max d_{ha}
DeepSWE	7.17984	100.00000	100.00000	2.19163	2.19163
LiveBench	0.16204	0.00000	33.33333	0.03270	0.03270
TUA	6.38229	83.33333	16.66667	1.04591	2.63586
Resolve	10.47904	0.00000	33.33333	2.89723	2.89723
AA	15.75598	66.66667	65.74074	3.50882	6.15204
MLE	3.15393	83.33333	23.80952	0.48617	0.83425
APEX	9.82736	66.66667	22.22222	2.42651	4.51895
CAR	11.53393	69.44444	9.83796	1.82431	4.61276
BFCL	27.92530	62.76923	69.16136	4.13128	9.36486
Banking	100.00000	100.00000	0.00000	100.00000	100.00000

Table 17: Maximum competition-rank spans and the matched weighted family ratio. Normalized spans are percentages; raw spans are places.

Panel	Raw span	Normalized span	SLAI / max d_{ha}
DeepSWE	15	41.66667	1.00000
LiveBench	5	8.92857	1.00000
TUA	8	40.00000	1.06947
Resolve	3	42.85714	1.00000
AA	19	61.29032	1.02067
MLE	9	33.33333	1.52197
APEX	3	42.85714	1.00000
CAR	6	54.54545	1.81405
BFCL	36	92.30769	2.98205
Banking	1	100.00000	1.00000

E.2 What the comparisons show in the saved panels

The full table exposes several distinct behaviors. LiveBench and Resolve preserve their champion sets while retaining nonzero LAI and SLAI. MLE's mean ordinary distance is below its largest fixed distance, and that largest distance is below its full-family LAI. BFCL makes this distinction particularly visible: mean $d = 22.12465\%$, maximum $d = 36.73077\%$, and $\mathrm{LAI} = 68.39744\%$. The

pair F10/F25 maximizes ordinary d. Compared with the full family, it leaves a total relation-width deficit of $19/60$, or 31.66667 percentage points, distributed over 269 configuration pairs. Some deficits are half a unit, so the count 269 is not obtained by multiplying 31.66667% by 780.

The weighted comparison uses a different baseline, $\max d_{ha} = 9.36486\%$, against SLAI $= 27.92649\%$. The 18.56163-point gap is area under the same position–amplitude weighting, and is not an increase in predictive accuracy. APEX supplies an equality case for both the unweighted family bound and the weighted one. Its inclusion shows that a nonzero family increment is not built into every multi-view panel.

BFCL's $1 - W_K$ and SLAI both round to 27.93% at two decimal places, but their unrounded values differ. Numerical proximity in one panel is different from an identity. Conversely, strong cross-panel correlation does not imply one statistic determines the other: the exact synthetic and CAR matches test determination directly. The comparison therefore combines algebraic reductions, descriptive agreement, and controlled non-identification instead of treating any single correlation as a sufficiency test.

E.3 Which information each summary retains

Table 18: Questions answered by the comparison measures. "Family" means the declared allowed views; it does not imply a probability distribution over them.

Summary	Information retained	Limitation exposed by the controls
Kendall / ordinary d	Pair-order disagreement for fixed endpoints	Adjacent reversals at different locations can coincide.
Spearman	Squared rank displacement for strict rankings	Same displacement vector magnitude can hide location.
$1 - W_K$	Concordance of rank sums across views	Does not specify a witness for each changed relation.
Winner / Top-5	Changes to a selected leading set	Relations outside that set may change while it stays fixed.
Maximum span	Largest individual rank movement	Does not describe the distribution of other changes.
Mean pair distance	Average discrepancy under entry weighting	Duplicate entries alter the implicit view-pair weights.
Maximum pair distance	Strongest one common endpoint comparison	Other endpoint pairs may support additional changes.
LAI	Full-family weak-relation width	Does not weight where or how far relations change.
SLAI	Full-family relation changes with joint position–amplitude witnesses	Depends on declared views and weights; is not a utility or quality score.

The duplication control has an important positive comparison: maximum distance, LAI, and a mean taken after weak-order deduplication are also invariant. Likewise, the simple first–second versus last-two swap is distinguished by winner change. The stronger second–third versus last-two construction was introduced to keep the winner fixed as well. These controls establish specific information differences while retaining cases in which conventional summaries already answer the intended question.

E.4 Direct comparisons with LLM sensitivity reports

We separate a sensitivity study's perturbation or reweighting procedure from the statistics used to report its results. Alzahrani et al. (2024), Section 4 and Figure 1, report normalized Kendall agreement $k_\tau = (1 + \tau)/2$ relative to the original ranking. Siska et al. (2024), Section 5.2 and Figures 2–4, report rank shifts, distributions of performance under weighting choices, and pairwise win frequencies. Hua et al. (2025), Section 2.3 and Table 1, use the mean Spearman correlation over all prompt pairs; Figure 2 reports performance means and standard deviations. These are direct LLM-evaluation precedents for the summaries examined here. We reuse their reporting objects on declared rankings, without claiming to rerun their prompt generation, random-weight sampling, or judges.

Table 19: Repeating the first view 50 times in every design panel. All values are percentages; deduplicated mean is unchanged. These copies are controls and are excluded from the main panel counts.

Panel	Mean before	Mean after	Dedup. mean	SLAI before	SLAI after
DeepSWE	15.24024	0.58616	15.24024	2.19163	2.19163
LiveBench	1.50376	0.05784	1.50376	0.03270	0.03270
TUA	10.35714	0.83375	10.35714	2.81897	2.81897
Resolve	19.64286	0.75549	19.64286	2.89723	2.89723
AA	19.22043	1.11355	19.22043	6.27919	6.27919
MLE	7.49559	0.60453	7.49559	1.26970	1.26970
APEX	14.28571	0.80863	14.28571	4.51895	4.51895
CAR	14.35185	3.09140	14.35185	8.36777	8.36777
BFCL	22.12465	11.26793	22.12465	27.92649	27.92649
Banking	100.00000	3.84615	100.00000	100.00000	100.00000

For the saved panels, Table 15 already reports mean $(1 - \tau_b)/2$ and $(1 - \rho)/2$ on identical inputs. Thus mean normalized Kendall agreement is one minus its Kendall column, and mean Spearman correlation is one minus twice its Spearman column. Average ranks and τ_b retain published ties explicitly. DeepSWE has normalized Kendall agreement 0.85717 and Spearman correlation 0.85641 while its leading configurations exchange first and twelfth place. MLE has mean Spearman correlation 0.95793 despite winner-set changes in five of its six view pairs. These numbers show why agreement and conditional winners should be reported together; they do not establish that SLAI alone determines a model choice.

For families with more than two views, our mean Kendall agreement averages all view pairs; this differs from retaining a separate agreement value against one original ranking as in Alzahrani et al. (2024). The control below matches the stated averages and marginals, not that full vector of reference-relative agreements.

Table 20: An exact synthetic control for sensitivity-report summaries. Both families contain four distinct strict rankings of the same four configurations. Frequencies are uniform over these four views; no model is run.

Reported object	Family X	Family Y
Rankings (best to worst)	ABCD, DCBA, BADC, CDAB	ABCD, DCBA, CADB, BDAC
Each configuration's rank distribution	Uniform on $\{1, 2, 3, 4\}$	Same
Each configuration's winner frequency	$1/4$	$1/4$
Every pair's win frequency	$1/2$	$1/2$
Mean Kendall τ / Spearman ρ	$-1/3$ / $-1/3$	Same
Maximum ordinary distance / LAI	1 / 1	Same
SLAI	$2/3$	$7/9$

Table 20 gives a stronger, exact information comparison. Every configuration occupies each rank once in both families. Consequently its complete rank distribution, rank mean $5/2$, rank variance $5/4$, and winner frequency $1/4$ coincide. Assigning the synthetic score $5 - r$ also makes each configuration's full score distribution, mean and variance coincide. Each ranking is accompanied by its reverse, so each of the six configuration pairs has win frequency $1/2$ in both families. Direct enumeration of the six view pairs gives mean Kendall $\tau = -1/3$ and mean Spearman $\rho = -1/3$ in both; maximum ordinary distance and LAI are both one. Nevertheless, applying Equations 3–5 gives $\mathrm{SLAI}(X) = 2/3$ and $\mathrm{SLAI}(Y) = 7/9$. Independent integer-cell integration verifies all twelve configuration-pair contributions. The reproduction script retains every rectangle and its source view pair.

For a direct arithmetic check, order the configuration pairs as AB, AC, AD, BC, BD, CD. Their integrated contributions are $(1/3, 2/3, 1, 1, 2/3, 1/3)$ in X and $(2/3, 2/3, 1, 1, 2/3, 2/3)$ in Y. Averaging each vector gives the stated SLAI. The difference resides in the joint arrangement of configurations across views, which the listed marginal and averaged summaries do not identify. A full ranking matrix retains this information and can be used to compute SLAI. We therefore make no claim that

sensitivity analysis as a whole cannot recover it. The example also does not invalidate the empirical finding of Hua et al. (2025) that a different evaluator can reduce sensitivity on their tasks. It isolates a distinction between reporting objects.

Two other LLM sensitivity measures address different inputs. POSIX measures changes in response log-likelihood under intent-aligned prompts (Chatterjee et al., 2024); ProSA's PromptSensiScore measures instance-level prompt sensitivity (Zhuo et al., 2024). A public ranking table lacks the conditional response probabilities or aligned per-instance prompt outcomes needed to compute these measures. We retain them as related work rather than substitute invented values or relabel an aggregate ranking statistic as their implementation.

E.5 A COMPARISON WITH POSITION-SENSITIVE RANKING MEASURES

This comparison bounds the novelty claim; it does not designate these measures as established LLM-evaluation baselines. Weighted rank measures already address the difference between leading and trailing disagreements. We use Vigna's additive hyperbolic weighted Kendall correlation τ_h (Vigna, 2015). An importance rank $R_i \geq 0$ gives item i weight $1/(R_i + 1)$ and pair i, j the sum of their weights. We use the symmetric convention that averages correlations based on decreasing lexicographic importance under $(s(q), s(q'))$ and $(s(q'), s(q))$, retaining the weighted tie normalization. Table 21 reports $(1 - \tau_h)/2$ on the same configuration rows and all 383 design-panel view pairs. The implementation is `scipy.stats.weightedtau` with `rank=True`, `additive=True`, and the default hyperbolic weigher; negative average ranks preserve the oriented weak orders exactly. Mean and maximum describe two explicit family aggregations, neither of which supplies the SLAI threshold surface merely by reporting its scalar.

Table 21: Position-sensitive reference comparison on the same saved design-panel rankings. Vigna's symmetric additive hyperbolic weighted Kendall distance is averaged or maximized over the declared view pairs. Percentages describe this measure's own scale, not improvement over another measure.

Panel	View pairs	Mean weighted Kendall	Max weighted Kendall
DeepSWE	1	23.29740	23.29740
LiveBench	1	0.93644	0.93644
TUA	6	9.00126	16.26848
Resolve	1	13.79520	13.79520
AA	3	20.45268	32.54669
MLE	6	6.61167	11.05323
APEX	3	14.45220	22.56430
CAR	36	14.24150	28.18942
Banking	1	100.00000	100.00000
BFCL	325	20.73260	43.91702

For complete strict lists on the same n items, rank-biased overlap (RBO) with persistence $p \in (0, 1)$ is

$$\mathrm{RBO}_p = (1 - p) \sum_{d=1}^{n} p^{d-1} \frac{|L_{1:d} \cap L'_{1:d}|}{d} + p^n. \tag{48}$$

The last term is exact here because both lists contain the same complete item set, so agreement is one at every deeper prefix. We report $1 - \mathrm{RBO}_{0.9}$ only on the strict synthetic controls; we do not silently break published ties to produce RBO results for the empirical panels.

If one adjacent pair at positions $k, k+1$ is swapped, all RBO prefixes agree except at depth k. Writing $H_n = \sum_{j=1}^{n} 1/j$, the two reference distances reduce to

$$\frac{1 - \tau_h}{2} = \frac{1/k + 1/(k+1)}{(n-1)H_n}, \qquad 1 - \mathrm{RBO}_p = \frac{(1-p)p^{k-1}}{k}. \tag{49}$$

For weighted Kendall, only the swapped pair is discordant; its additive weight is $1/k + 1/(k+1)$ in either importance ordering, and the sum of all pair weights is $(n-1)H_n$. For Figure 6a, $n = 6$ and $k = 2$ or 5. Weighted Kendall gives $10/147$ versus $22/735$; RBO at $p = 9/10$ gives $9/200$ versus $6561/500000$; SLAI gives $4/375$ versus $1/375$. Thus all three distinguish these positions.

Exact rational enumeration checks these formulas at $n = 6, 10, 100$, and an independent pairwise implementation checks weighted Kendall against the library on all 120 strict five-item permutations.

This positive comparison limits the interpretation of Theorem 1: the theorem separates SLAI from its stated conventional-summary vector, not from all rank measures. Likewise, duplication invariance is shared by LAI, maximum distances, and means after deduplication. SLAI combines these properties with a different reporting object: a full-family surface whose position and amplitude conditions have an actual common view-pair witness. Other measures remain appropriate for their own comparison questions, and these calculations do not establish an advantage in predictive accuracy or decision utility.

E.6 Descriptive associations

For each design panel, we compare SLAI with mean normalized Kendall distance $(1 - \tau_b)/2$, mean normalized Spearman distance $(1 - \rho)/2$, tie-corrected Kendall inconsistency $1 - W_K$, normalized maximum competition-rank span, and largest ordinary and weighted view-pair distances. Spearman correlation uses average ranks. Competition rank is one plus the number of strictly better configurations and is used for reported rank spans; SLAI uses average ranks throughout.

Table 22: Cross-panel Spearman correlations with SLAI. These describe the saved heterogeneous panels, without population or significance inference.

Summary	All 10	Without Banking	Also without BFCL
Mean Kendall distance	.782	.700	.571
Mean Spearman distance	.842	.783	.690
$1 - W_K$	.964	.950	.929
Maximum normalized rank span	.973	.962	.946
Maximum ordinary pair distance	.988	.983	.976
Maximum weighted pair distance	.988	.983	.976

The positive associations show shared sensitivity to ranking changes. Their strength also limits an empirical claim that SLAI radically reorders these panels. Its added information here concerns the magnitude, position, and family coverage of variation. Exact summary collisions provide a different test: whether those summaries determine the SLAI value.

E.7 Exact matching in real data

The bounded search considers all 383 real view pairs in the design panels. Comparisons are made within each panel's full configuration set. The predeclared match vector comprises ordinary preference distance, Kendall correlation, Spearman correlation, tie-corrected Kendall inconsistency, maximum competition-rank span, and winner-set Jaccard distance. Signed rational-square representations match correlations exactly at the saved score precision.

One complete-vector match is found in CAR. Let P_A compare Base/single with Hallucination/all-three, and P_B compare Base/any-three with Hallucination/all-three. Both use all 12 configurations and 66 model pairs. The difference is $49/21296$, or approximately 0.23009 percentage points. The comparisons share an endpoint, are discovered in a finite saved-data search, and do not hold all positional information fixed. An additional 2,312 matches on maximum span alone are retained but do not count as full-vector collisions. The matched set excludes maximum weighted distance, which equals SLAI for a two-view family. The source check covers 36 score cells and independently checks the 132 model-pair contributions.

Table 23: CAR's exact summary collision. The six matched quantities agree, while the two-view SLAI values differ. These values are distinct from CAR's nine-view SLAI.

Quantity	P_A	P_B
Ordinary preference distance / LAI	$2/11$	$2/11$
Kendall τ_b	$\sqrt{147/352}$	$\sqrt{147/352}$
Spearman ρ	$\sqrt{56169/81224}$	$\sqrt{56169/81224}$
Tie-corrected $1 - W_K$	$8/95$	$8/95$
Maximum competition-rank span	4	4
Winner-set Jaccard distance	1	1
Two-view SLAI	$1411/63888$	$79/3993$
SLAI (%)	2.20855	1.97846

Table 24: The 36 saved CAR score cells used by the two matched comparisons. Values are percentages at source precision; all 12 configuration rows are retained.

Configuration	Base single	Base any 3	Hall. all 3
Claude-Opus-4.6 (thinking)	84	93	48
Claude-Opus-4.5 (thinking)	77	86	52
GPT-5 (thinking)	76	88	60
Claude-Sonnet-4 (thinking)	74	83	46
GPT-5.2 (thinking)	74	85	57
Gemini-2.5-flash (thinking)	67	80	41
Gemini-2.5-pro (auto-thinking)	67	80	34
GPT-4.1	64	69	31
Qwen3-32b (thinking)	62	77	27
Gemini-2.5-flash	53	62	32
GPT-Oss-120b (thinking)	39	42	37
xLAM-2-32b	38	42	11

F Weight and Tie Sensitivity

F.1 Position and amplitude preferences

The tested family is

$$M_{\alpha,\beta} = \int_{[0,1]^2} \mathcal{A}(u,v)\, dF_\alpha(u)\, dF_\beta(v), \qquad F_\alpha(u) = u^\alpha \quad (\alpha > 0). \tag{50}$$

For $\alpha = 0$, the measure is a point mass at zero; likewise for $\beta = 0$. Thus $M_{0,0} =$ LAI and $M_{1,1} =$ SLAI. Positive weights can be evaluated by mapping rectangle corners to (h^α, a^β). A zero exponent means evaluating that threshold at zero, including boundary cases, rather than relying on an ambiguous floating-point 0^0 operation.

The grid $\alpha, \beta \in \{0, 0.5, 1, 2\}$ gives 16 settings. Larger positive exponents emphasize stricter thresholds; zero removes the corresponding threshold requirement. Across all ten panels, the minimum Spearman correlations with default SLAI are 0.964 for the nine positive settings and 0.952 across all settings. Without Banking they are 0.950 and 0.933. Among the nine non-Banking panels, positive settings change at most two of 36 panel-pair orders and move any panel by at most two positions. The defaults were not optimized on these panels. Arbitrary location functions and nonseparable threshold measures are outside the tested family.

F.2 Tie treatment

The 25 settings comprise the default, four deterministic variants, and 20 shared random priorities. Two deterministic variants replace average-rank coordinates with competition ranks or the last occupied tied position while keeping weak-order preferences. The other two retain average ranks but assign tie-only relation changes weights $1/4$ or $3/4$. Strict reversals remain 1, and unchanged relations remain 0. These penalties are symmetric under swapping model identities.

Each random setting uses one total configuration priority per panel, shared across all its views. Thus identical weak rankings and their duplicates share tie-breaking randomness. The saved seeds run from 2026091900 through 2026091919. The priorities are assumptions about hidden strict orders, not observations of those orders or confidence intervals. The source data and all seeds are retained.

Table 25: Tie sensitivity on the ten design panels. Entries are SLAI percentages; random ranges cover the 20 shared-priority settings.

Panel	Default	Competition	Last tied	Tie penalty $1/4, 3/4$	Random range
DeepSWE	2.19163	2.18358	2.20359	2.05897, 2.32429	1.97588–2.49196
LiveBench	.03270	.03270	.03270	.03270, .03270	.03270–.03270
TUA	2.81897	2.89673	2.77857	2.62653, 3.01142	2.43095–3.43393
Resolve	2.89723	3.06122	2.76968	2.71501, 3.07945	2.18659–3.82653
AA	6.27919	6.27549	6.29217	6.17378, 6.38460	6.05573–6.48780
MLE	1.26970	1.37365	1.18576	1.17996, 1.35944	1.11028–1.49132
APEX	4.51895	4.51895	4.51895	4.51895, 4.51895	4.51895–4.51895
CAR	8.36777	8.09855	9.02517	7.87159, 8.86395	6.64914–9.91109
BFCL	27.92649	28.49667	27.56903	27.72227, 28.13071	27.62854–28.60184
Banking	100	100	100	100, 100	100–100

Excluding Banking, deterministic variants have cross-panel correlation at least 0.983333 with the default and random variants at least 0.950000. BFCL, CAR, AA, and APEX retain that order and LiveBench remains lowest throughout. Local panel ordering and individual values change, especially for Resolve, TUA, and CAR. Banking has no ties and only two configurations, so its invariance adds little evidence about tie handling. The additional four source panels have not been assigned this entire 25-policy robustness claim.

F.3 A MATCHED COMPARISON FOR THE BFCL FAMILY GAP

For each tie policy, the strongest fixed-pair distance is recomputed using that same policy. The default factor is 2.98205; competition ranks give 2.99563, last tied positions 2.98997, and the $1/4$ and $3/4$ penalties give 2.99279 and 2.97155. The per-seed random ratios range from 2.92076 to 3.05204. Numerator and denominator ranges must not be cross-divided because their extrema can occur at different seeds. The conclusion is a persistent family-coverage gap under these declared policies, without universality over all tie rules.

F.4 THE FULL WEIGHT GRID AND THE ONE-FACTOR ABLATIONS

Tables 26 and 27 report all 160 panel–weight values. The zero-coordinate rows make clear how much of a value remains after removing location, amplitude, or both. Because higher exponents place more mass on stricter thresholds, values decrease coordinatewise as established in Proposition 5. This numerical decrease is distinct from cross-panel order stability.

Table 26: All 16 threshold-weight settings, part 1. Entries are percentages. Zero removes that threshold; (1,1) is default SLAI.

α	β	DeepSWE	LiveBench	TUA	Resolve	AA
0	0	15.24024	1.50376	20.00000	19.64286	28.83065
0	0.5	5.97604	0.28248	8.76535	9.27736	14.05308
0	1	2.51606	0.05594	3.98214	4.46429	7.29058
0	2	0.52103	0.00253	0.89970	1.08874	2.25038
0.5	0	13.80584	1.12472	16.55307	15.51102	25.73523
0.5	0.5	5.48670	0.21240	7.27731	7.37319	12.78850
0.5	1	2.33581	0.04228	3.31892	3.56306	6.72709
0.5	2	0.49246	0.00193	0.75630	0.87084	2.11292
1	0	12.70333	0.87160	14.02976	12.50000	23.52855
1	0.5	5.10198	0.16437	6.17199	5.97500	11.83216
1	1	2.19163	0.03270	2.81897	2.89723	6.27919

α	β	DeepSWE	LiveBench	TUA	Resolve	AA
1	2	0.46899	0.00149	0.64525	0.70869	1.99578
2	0	11.13974	0.56341	10.67604	8.60058	20.44837
2	0.5	4.54294	0.10475	4.67731	4.14502	10.41772
2	1	1.97788	0.02058	2.13101	2.01869	5.58569
2	2	0.43294	0.00092	0.48748	0.49296	1.80264

Table 27: All 16 threshold-weight settings, part 2. Entries are percentages. Zero removes that threshold; (1,1) is default SLAI.

α	β	MLE	APEX	CAR	BFCL	Banking
0	0	14.28571	21.42857	37.87879	68.39744	100.00000
0	0.5	5.06829	10.83406	19.95869	44.99698	100.00000
0	1	1.88247	5.61224	10.70937	30.26668	100.00000
0	2	0.29010	1.60350	3.24631	14.46723	100.00000
0.5	0	11.61367	17.80548	33.08488	64.77166	100.00000
0.5	0.5	4.13151	9.29094	17.45167	42.91815	100.00000
0.5	1	1.53184	4.93778	9.37635	29.00868	100.00000
0.5	2	0.23411	1.46272	2.85329	13.95066	100.00000
1	0	9.73692	15.81633	29.47658	61.88774	100.00000
1	0.5	3.44616	8.39087	15.55905	41.18756	100.00000
1	1	1.26970	4.51895	8.36777	27.92649	100.00000
1	2	0.19149	1.36401	2.55497	13.48787	100.00000
2	0	7.29069	13.62974	24.61808	57.41652	100.00000
2	0.5	2.53410	7.32559	12.99875	38.38374	100.00000
2	1	0.91680	3.98792	6.99861	26.11705	100.00000
2	2	0.13350	1.22345	2.14812	12.68297	100.00000

For BFCL, the four corner ablations $(0,0),(1,0),(0,1),(1,1)$ give 68.39744%, 61.88774%, 30.26668%, and 27.92649%. Both weighting dimensions reduce the unweighted family width, with amplitude producing the larger reduction in this panel. The same calculation need not preserve a fixed relative effect in another panel. For DeepSWE the four corresponding values are 15.24024%, 12.70333%, 2.51606%, and 2.19163%.

Table 28 gives every setting's panel-order comparison. These correlations describe an ordering of measured sensitivity among the selected panels, rather than model rankings within an individual benchmark. Reporting both levels separately avoids treating a stable cross-panel ordering as proof that every configuration comparison is stable.

Table 28: Panel-order sensitivity for each setting. The last two columns exclude Banking and give maximum displacement and changed pair orders among nine panels.

α	β	ρ, all 10	ρ, without Banking	Max move	Flips
0	0	0.98788	0.98333	1	1
0	0.5	1.00000	1.00000	0	0
0	1	1.00000	1.00000	0	0
0	2	1.00000	1.00000	0	0
0.5	0	0.98788	0.98333	1	1
0.5	0.5	1.00000	1.00000	0	0
0.5	1	1.00000	1.00000	0	0
0.5	2	1.00000	1.00000	0	0
1	0	0.96364	0.95000	2	2
1	0.5	0.98788	0.98333	1	1
1	1	1.00000	1.00000	0	0
1	2	1.00000	1.00000	0	0
2	0	0.95152	0.93333	2	3
2	0.5	0.96364	0.95000	2	2
2	1	0.98788	0.98333	1	1

α	β	ρ, all 10	ρ, without Banking	Max move	Flips
2	2	1.00000	1.00000	0	0

F.5 Full matched BFCL tie results

Table 29 lists all 25 settings rather than only the smallest and largest observed values. In each row the fixed-pair denominator is recomputed under the same tie rule as the full-family numerator. The persistence of a ratio near three is consequently not an artifact of retaining the default denominator.

Table 29: BFCL full-family and strongest fixed-pair values under each matched tie policy. Both values are percentages; ratios use unrounded values in the same row. Random labels give the final two digits of seeds 2026091900–2026091919.

Policy	SLAI	Max d_{ha}	Ratio
Default	27.92649	9.36486	2.98205
Competition	28.49667	9.51274	2.99563
Last tied	27.56903	9.22051	2.98997
Tie penalty $1/4$	27.72227	9.26303	2.99279
Tie penalty $3/4$	28.13071	9.46669	2.97155
Random 00	28.34332	9.28670	3.05204
Random 01	27.88424	9.38755	2.97034
Random 02	28.00439	9.41393	2.97478
Random 03	28.36456	9.41520	3.01264
Random 04	28.00616	9.37482	2.98738
Random 05	27.88609	9.32454	2.99061
Random 06	28.32360	9.31165	3.04174
Random 07	27.86181	9.47483	2.94061
Random 08	28.46799	9.43066	3.01866
Random 09	27.62854	9.37077	2.94837
Random 10	28.20049	9.43222	2.98980
Random 11	27.84116	9.53215	2.92076
Random 12	28.01636	9.35539	2.99468
Random 13	28.10251	9.38312	2.99501
Random 14	28.24795	9.40049	3.00495
Random 15	28.60184	9.41874	3.03670
Random 16	27.91117	9.34713	2.98607
Random 17	28.06091	9.43715	2.97345
Random 18	28.15388	9.46101	2.97578
Random 19	27.88120	9.29959	2.99811

Shared random priorities preserve duplicate-ranking invariance because identical tied groups receive identical tie resolution in every copy. Redrawing a separate priority for each column would introduce apparent disagreement between duplicated evidence and test a different perturbation. Relabeling configurations is symmetric in the random-priority distribution, but regenerating a fixed seed after changing the order of identifiers need not give the same realized priority. Neither the twenty-seed range nor the five deterministic policies estimates a sampling confidence interval.

F.6 What the sensitivity checks jointly support

The weighted-family gap in BFCL persists across the declared tie policies, and the leading cross-panel order BFCL, CAR, AA, APEX persists across the tested positive weight grid and tie settings. These are two separate robustness statements: one concerns a within-panel family-to-fixed-pair comparison, and the other concerns between-panel ordering. Resolve, TUA, and DeepSWE can exchange nearby positions, and individual SLAI values can vary more than the panel order suggests. The full tables retain those changes.

The analysis does not perturb hidden decimals, estimate alternative source measurements, or average over hypothetical view distributions. Those operations would require an additional measurement or uncertainty model. Here the saved score matrix is fixed and only the declared threshold and tie

conventions change. The four added main sources retain their frozen default calculations; they are not silently included in the design panels' complete 25-policy tie study.

G Evidence Removal and Reproducibility

Ranking geometry and task-level interpretation require different evidence. The inherited grouping-removal control makes that distinction explicit: it retains the design and success totals, but removes task identities and all derived statistics that would recover exact coverage. The constructions below show which answer remains determined. They use the original 37-configuration research materials; the older claim-identification study's 24 requests and independent review are not counted as additional SLAI validation.

G.1 A constructive proof of the missing-grouping bounds

Suppose only S successes from T tasks and K attempts per task remain. If C tasks are covered, each covered task contributes at least one and at most K successes. Therefore

$$\left\lceil \frac{S}{K} \right\rceil \leq C \leq \min(S, T). \tag{51}$$

Both endpoints are attainable. For the lower endpoint, put K successes on as many tasks as possible and the remainder, if any, on one more task. For the upper endpoint, distribute one success to each of $\min(S, T)$ tasks, then distribute the remaining successes without exceeding capacity K. More generally, every integer C in the interval is feasible: start with one success on each of C tasks and place the remaining $S - C$ successes in the $C(K - 1)$ remaining slots.

In the DeepSWE example, $T = 113$, $K = 4$, and the totals are 335 and 291. The following constructions preserve those totals exactly: The reversal under compatible completions proves that

Table 30: Compatible task-success histograms when task grouping is removed. Each tuple lists counts at $c = 0, 1, 2, 3, 4$; these are mathematical completions, not additional observed outcomes.

Completion	Astra histogram / coverage	GPT-5.5 histogram / coverage
GPT-5.5 covers more	$(29, 0, 0, 1, 83)$ / 84	$(0, 53, 1, 0, 59)$ / 113
Astra covers more	$(0, 39, 0, 0, 74)$ / 113	$(40, 0, 0, 1, 72)$ / 73
Observed grouping	$(22, 2, 8, 7, 74)$ / 91	$(11, 20, 17, 23, 42)$ / 102

coverage order is not determined by the success totals. The single-attempt order remains determined because its denominator and totals are retained. Once actual task identities are restored, coverage is again 91 versus 102. Permuting intact rows leaves every statistic unchanged; deleting grouping changes the available information. A totals-only ablation must remove not only explicit task IDs but also histograms, coverage totals, or other fields that encode the same answer.

G.2 Why format marginals do not identify joint task success

BFCL's marginal format scores suffice for the saved ranking analysis. They do not identify the set of tasks on which a configuration succeeds under every format. Even with a known common support and exact marginal counts, intersections can remain undetermined. For example, on four tasks, two formats can each have two successes. If their success sets are identical, both-format success is two; if their sets are disjoint, it is zero. The marginal score vector is identical in the two cases.

With m binary events on a genuinely common support of size T, marginal counts c_q imply the generic intersection bounds

$$\max\left(0, \sum_q c_q - (m-1)T\right) \;\leq\; \left|\bigcap_q S_q\right| \;\leq\; \min_q c_q. \tag{52}$$

The upper bound is containment in each success set; the lower bound follows by applying the union bound to their complements. Applying these bounds to a particular source requires a verified

common support and counts with the required precision. We do not replace the missing BFCL task-by-format matrix with a joint-success estimate or claim an empirical bound from rounded percentages. The ranking witnesses and the unresolved joint-outcome question can therefore coexist without contradiction.

G.3 From saved source to reported number

The reproducibility path has three stages. First, source records are mapped to an explicitly named configuration and view, preserving the source's support and score precision. Second, every column is ranked on that fixed row set, producing preferences, relation classifications, and doubled average ranks. Third, view-pair witnesses are integrated into pair contributions, averaged into panel totals, and retained for inspection. A table of final SLAI values alone would omit the second and third stages needed to diagnose an implementation or scope error.

Table 31: Reproduction map from source records to reported results. Paths are relative to the reproduction materials' Paper 2 directory.

Object	Input and output responsibility
Source records	Archived inputs include DeepSWE labels, TUA rewards, Resolve counts, MLE reports, and published aggregate scores.
Score and target registry	`output/main_figures_20260918_v5/inputs/` contains `all_scores_ranks.csv`, `target_registry.json`, and the BFCL format key.
Default exact integration	`output/compact_atlas_20260918_v4/pareto_sweep.py` computes pair contributions, frontiers, and panel totals.
Metric comparison	The same directory's `data/neighbor_summary.csv` binds conventional measures, moments, and fixed-pair comparisons.
Independent grid	`output/slai_argument_audit_20260919/` retains the direct-grid checks and cross-panel comparisons.
Tie and first extension	`output/slai_validation_extension_20260919/` retains protocols, source mappings, 25-policy results, and CAR matches.
Second extension	`output/slai_preregistered_batch2_20260919/` retains admission decisions, complete and auxiliary panels, and source-cell checks.
Synthetic controls	`output/slai_adversarial_controls_20260919/` retains theorem statements, finite checks, and explicit ranking constructions.

Verification levels. Exact reaggregation checks that the saved labels or sufficient statistics produce the reported scores. Source-cell checks bind those inputs to the saved release. Independent integration checks the rank geometry against a different algorithm. Mathematical proofs establish claims beyond the finite panels. None of these operations reruns a model or independently rejudges its original output. Their counts are therefore reported as checks or computation operations, separately from task counts and source counts.