CELLS OR VARIANT FAMILIES? RESAMPLING-UNIT SENSITIVITY IN REPEATED-VARIANT VLM BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated-variant vision–language model (VLM) benchmarks score multiple prompts, option rotations, or counterfactual images derived from a shared source. Treating those cells as independent can change reported uncertainty while leaving the observed score, eligible cells, and weights unchanged. We conduct a version-pinned descriptive audit of 23 public response matrices from DynaMath, MMBench V11, and VB, comparing cell-IID and source-family percentile-bootstrap intervals. Median family-to-cell interval-width ratios are 2.329, 1.827, and 0.865, respectively; the MMBench result is a rotated-prompt cell diagnostic, not its official CircularEval score. Across ten fixed model pairs, one unadjusted pointwise comparison changes from positive to unresolved; the other nine are unchanged. A post-hoc Bonferroni analysis also changes one comparison, but a different one. We report a scorer correction to our earlier analysis and present a descriptive re-analysis rather than a prospectively verified confirmatory study. The contribution is a bounded VLM-specific cross-benchmark measurement, not a new bootstrap method or a field-wide prevalence claim. The results motivate declaring the target population, scoring and eligibility rules, and resampling unit, and testing sensitivity while keeping the observed statistic fixed.

1 INTRODUCTION

Modern VLM evaluations often expand one source into many scored cells. DynaMath renders ten parameterized variants of each seed problem Zou et al. (2025); MMBench cyclically rotates multiple-choice options within a question Liu et al. (2024); and VB applies controlled visibility and perspective transformations within a source scene Tripathi (2026). These designs are valuable because they expose failures hidden by a single static prompt. They also create an inferential choice: is the population represented by all derived cells, or by independently sampled source families whose cells remain together?

Treating repeated measurements as independent is the classical pseudoreplication problem Hurlbert (1984). In benchmark analysis, IID and clustered bootstraps can differ in interval calibration and decision error Bakshy and Frachtenberg (2015). Cluster-aware inference is not new to multimodal evaluation: recent work resamples images Liu et al. (2026), episodes de Curto et al. (2026), or templates Chen et al. (2026), and counterfactual chart evaluation performs family-level paired tests Jiang et al. (2026). More directly, Neuhof and Benjamini vary the inferential unit across subjects, questions, and prompt variants in PromptEval/MMLU while holding the paired score numerator fixed, then propagate pairwise tests to rank intervals Neuhof and Benjamini (2026a). For source-derived VLM variants, we ask how much interval widths change between cell and family resampling across released benchmark formats, and whether that choice changes any fixed paired-model conclusion.

We analyze 23 public response matrices from three benchmarks and ten model pairs specified by a documented architecture/capacity or complete-availability rule. The comparison holds the observed score, eligible cells, and cell weights fixed while changing the resampling unit. All scores come from released outputs, so the analysis requires no model training or API calls. We present a corrected descriptive re-analysis: our earlier VB scorer was amended after review, and conflicting local

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.

timestamps prevent independent verification of prospective status. The scorer and chronology are documented in Sections 3 and 6 and Appendix G. We make three contributions:

- We reconstruct source families in three released response-matrix formats and compare cell-IID and family-cluster 95% intervals for the identical observed cell-weighted statistic.
- We quantify materiality with a frozen width-ratio equivalence band and ten fixed pointwise paired comparisons, then add a clearly labeled post-hoc Bonferroni-adjusted sensitivity.
- We provide a reproducible audit trail: pinned upstream commits and hashes, deterministic resampling, complete tables, two diagnostic seeds, preserved instrumentation failures, and explicit provenance and scorer corrections.

The direction of the interval-width change depends on the benchmark. Family resampling more than doubles interval width in DynaMath and increases it by roughly 83% in our MMBench diagnostic, yet narrows intervals in VB. One of ten pointwise labels changes; one also changes under nominal Bonferroni-adjusted intervals, although it is a different pair. These opposite directions motivate declaring the source unit and measuring sensitivity to that choice.

2 Related Work and Novelty Boundary

Repeated-variant VLM benchmarks. DynaMath groups ten dynamically generated visual math variants under each of 501 seeds and reports average, worst-case, and consistency-oriented metrics Zou et al. (2025). MMBench's CircularEval requires consistency over cyclic option rotations, already recognizing the question family as a meaningful scoring object Liu et al. (2024). VB crosses controlled image and question transformations within 100 source families Tripathi (2026). Other recent benchmarks similarly organize counterfactuals or matched images around common sources Jiang et al. (2026); Chen et al. (2026). These works motivate family-aware evaluation but do not report the fixed-statistic cell-IID versus family-cluster interval comparison and its effect on documented fixed model gaps.

Uncertainty units and leaderboard conclusions. Pseudoreplication and cluster resampling are established statistical issues Hurlbert (1984); Bakshy and Frachtenberg (2015). Aggregate benchmark uncertainty has been studied with bootstrap and hierarchical models Longjohn et al. (2025), and recent rank-interval work makes task variability and multiplicity explicit Neuhof and Benjamini (2026b;a). The closest current neighbor, Neuhof and Benjamini (2026a), already shows on PromptEval/MMLU that changing whether subjects, questions, or prompts are inferential units changes the resulting score and rank uncertainty. In VLM research, WEP keeps questions sharing an image together in a paired image-cluster bootstrap across four benchmarks Liu et al. (2026); REFLECT resamples counterfactual templates Chen et al. (2026); zero-shot VLM control resamples complete episodes de Curto et al. (2026); and Chartographer uses family-indexed sign-flip tests across three chart datasets Jiang et al. (2026). Separately, changing an annotator pool can leave a small observed leaderboard unchanged while bootstrap analysis reveals substantial displacement risk Jha (2026). A newer near-tie audit reweights observed LLM benchmark items toward source- and easiness-balanced low-DIF subsets and uses source-block bootstrap diagnostics, finding excess ordering reversals among sub-one-point cross-family pairs in four of five benchmarks Zheng and Yang (2026). That intervention changes benchmark composition and weights, whereas our comparison holds the observed score, weights, and eligible cells fixed. Accordingly, we do *not* claim a new bootstrap method, the first unit-aware model comparison, or a universally correct uncertainty estimator. Our remaining contribution is a descriptive re-analysis of released outputs, including a correction to our own earlier scorer: a version-pinned table that holds the observed cell-weighted statistic fixed, contrasts source-derived cells with their original families across three VLM benchmark designs, and reports both interval-width directions and all ten fixed paired decisions.

3 AUDIT DESIGN

3.1 OBSERVED STATISTIC AND UNCERTAINTY CONTRACTS

Let source family $g \in \{1, \ldots, G\}$ contain m_g eligible cells with scores $y_{gj} \in [0, 1]$. Our observed statistic is the pooled cell-weighted mean

$$\widehat{\mu} = \frac{\sum_{g=1}^{G} \sum_{j=1}^{m_g} y_{gj}}{\sum_{g=1}^{G} m_g}. \tag{1}$$

We hold this scoring function, its cell weights, and the eligible cells fixed. Only the bootstrap unit changes. The cell-IID bootstrap samples individual cells with replacement. The family-cluster bootstrap samples G source families with replacement, retains every eligible cell in a sampled family, and recomputes the pooled mean. With constant m_g, this is equivalent to resampling family means; with variable m_g, it recomputes the same pooled statistic, rather than an unweighted mean of family means.

Target interpretation. For the fixed released corpus, the displayed point score has no sampling uncertainty. The intervals instead represent two hypothetical superpopulation contracts: future eligible cells sampled independently, or future source families sampled with their observed nested variants. We compare their empirical consequences; we do not establish that either bootstrap has optimal coverage for an unspecified deployment population. A family interval is meaningful only insofar as the released families represent the sources to which one intends to generalize. Keeping the observed statistic fixed does not by itself identify a common parameter of these hypothetical populations. In particular, VB scores condition on model-specific parse eligibility; we neither assume missingness is random nor recover performance on unscored cells.

For each model, 20,000 deterministic bootstrap replicates produce a 95% percentile interval. We summarize sensitivity as

$$R = \frac{q_{.975}^{\text{cluster}} - q_{.025}^{\text{cluster}}}{q_{.975}^{\text{IID}} - q_{.025}^{\text{IID}}}. \tag{2}$$

The original protocol uses a $[0.90, 1.10]$ band as a practical-equivalence convention. Ratios above 1.10 indicate wider family intervals; ratios below 0.90 indicate narrower ones. We retain that rule unchanged in the corrective analysis without claiming that local timestamps independently prove its prospective status.

For a fixed pair (A, B), we align identical family-cell keys and analyze $d_{gj} = y_{gj}^{A} - y_{gj}^{B}$. A 95% interval is labeled *positive*, *negative*, or *unresolved* depending on whether it lies above zero, below zero, or contains zero. A decision changes only if the label differs between contracts. Common resampling draws preserve pairing.

3.2 FROZEN BENCHMARK SNAPSHOTS

Table 1 describes the primary matrices. Repository commits, selected files, and bytes are SHA-256 bound in the prepared local audit trail. DynaMath and VB inclusion is exhaustive within their pinned public snapshots. The documented MMBench selection rule specifies three same-architecture capacity pairs rather than every upstream model. The ten-pair rule uses architecture lineage and capacity (four DynaMath, three MMBench) or complete pair availability (three VB). These content-based rules are recorded, but the timestamp limitation below prevents an independently verified claim of historical outcome blinding.

DynaMath. The pinned official release provides fourteen complete 501×10 correctness matrices. Family IDs are the seed-question identifiers and all ten generated variants remain nested within a seed. Categorical tokens are decoded exactly as released. No model or cell is missing.

MMBench V11. Family identity is recovered from the released index used by CircularEval. CircularEval itself is a useful positive control for family-aware scoring; our auxiliary question is what happens if its released rotations are treated as cells for uncertainty. The primary subset contains the 2,688 families with four aligned option rotations in all six selected model files. We score each

Table 1: Frozen primary matrices. Cells are per model. MMBench is a cell-level option-accuracy diagnostic, not the official CircularEval headline.

Benchmark	Models	Families	Cells/family	Cells	Analysis status
DynaMath	14	501	10	5,010	corrected descriptive
MMBench V11	6	2,688	4	10,752	descriptive diagnostic
VB	3	100	1–3	251–300	scorer-corrected descriptive

rotation using a pinned, no-API VLMEvalKit exact matcher Duan et al. (2024): an option token, the verbose form "answer is X," or exact option text is matched; unresolved predictions become `Z` and therefore incorrect. The primary subset has 271 fallbacks among 64,512 model-cells; all released rows have 302 among 73,050 (12,175 per file), with none silently dropped. The additional 529 families contain 1,423 rows per file. Appendix C separates the per-model counts. Crucially, this is a *cell-level rotated-prompt accuracy diagnostic*; it neither reproduces nor corrects MMBench's strict circular-family headline score.

VB. We use the three model output CSVs available at the pinned public commit and the repository's confidence-aware accuracy with abstention (CAA) semantics, including 0.25 credit for abstention. The official scorer retains every nonempty normalized label; a noncanonical label is an observed wrong prediction with zero credit, not a missing cell. A separate AI-assisted review found that A4 instead dropped three `INHERENTLY_NONVISUAL` InternVL3 cells. A5 preserves A4 and corrects that parser, yielding available counts of 300, 300, and 251 cells. Family bootstrap samples the 100 sources while retaining their 1–3 officially eligible cells. Qwen has 49 unscored headline cells: 1 BASE, 36 TEXT_FLIP, and 12 IMAGE_FLIP; only 59 of its families have all three cells. Each model pair uses its own common eligible cells: 300 for Gemma–InternVL and 251 for either Qwen pair. These differences need not equal differences of marginal model scores or form a transitive ranking on one shared support. All VB results are post-review scorer-corrected descriptive evidence.

3.3 Frozen decisions, robustness, and audit trail

The archived protocol contains two descriptive criteria and one implementation invariant: (H1) at least two benchmark median ratios lie outside $[0.90, 1.10]$; (H2) at least one of ten fixed pair labels changes; and (H3) changing only the resampling unit preserves the observed point estimate up to 10^{-12}. The protocol also retained the falsifying outcomes, including zero changed pairs.

Two additional seeds with 5,000 replicates test interval-band and pair-label stability, and A5 persists their interval endpoints. DynaMath drops each of ten variant positions in turn; MMBench repeats the analysis on all 3,217 variable-size families; and VB repeats a complete-three-cell sensitivity on each model's own eligible families. A declared fourth-cell sensitivity adds DOUBLE_FLIP to the three headline relations; it changes the evaluated score support and is kept separate from the primary analysis (Appendix D). Four stopped runs are preserved rather than overwritten: two exposed DynaMath categorical encodings, one exposed MMBench free-form predictions, and one exposed VB availability below the original threshold. The repair files and run outputs remain hash-bound, but inconsistent editable timestamp fields prevent an independently verifiable claim that every repair preceded its successor. A5 therefore treats the resulting numerical pattern as descriptive.

4 Results

4.1 Interval width changes in both directions

All three benchmark medians fall outside the original equivalence band (Table 2; Figure 1A). DynaMath family intervals are 2.198–2.508 times as wide as cell-IID intervals across all fourteen models, with median 2.329. MMBench ratios are 1.806–1.873 across all six models, with median 1.827. VB reverses direction: ratios are 0.846–0.947, with median 0.865; two models fall below 0.90 and one lies inside the equivalence band. The originally specified H1 rule is therefore satisfied descriptively.

Table 2: Primary family/cell 95% interval-width ratios. Counts partition models using the frozen equivalence band.

		Width ratio R		Model counts		
Benchmark	Models	Median	Range	< 0.90	$[0.90, 1.10]$	> 1.10
DynaMath	14	2.329	2.198–2.508	0	0	14
MMBench V11	6	1.827	1.806–1.873	0	0	6
VB	3	0.865	0.846–0.947	2	1	0

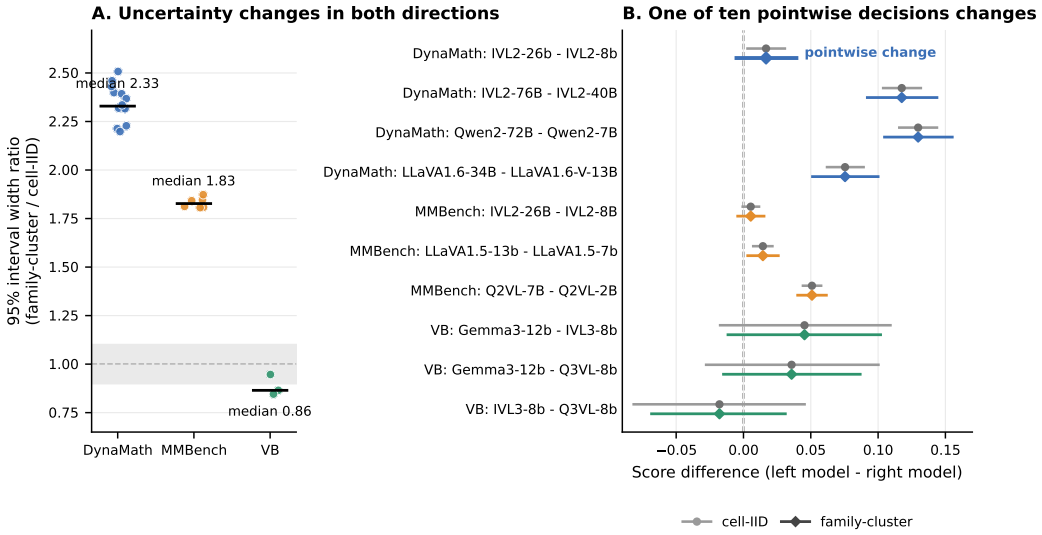

Figure 1: **Changing the resampling unit changes precision, but not always in the same direction.** (A) Points are models, black bars are benchmark medians, and the gray region is the frozen $[0.90, 1.10]$ equivalence band. (B) Pointwise cell-IID intervals (gray circles) and family-cluster intervals (colored diamonds) for ten fixed pairs. The highlighted DynaMath comparison is the only zero-containment decision change at the pointwise 95% level. DynaMath/MMBench gaps are accuracy differences; VB gaps are CAA differences on pair-specific supports.

The observed point estimates are identical under both contracts up to a maximum numerical difference of 2.22×10^{-16}, verifying the H3 implementation invariant. This is an uncertainty audit, not a correction to published point scores. The opposite directions are also substantively informative: retaining families can expose positive within-family dependence, as in DynaMath and MMBench, or a pattern compatible with within-family counterbalancing, as in VB. The word "cluster" alone does not determine which interval is wider.

4.2 One of ten fixed pointwise model-gap decisions changes

The DynaMath InternVL2-26B minus InternVL2-8B point gap is 0.01677. Its cell-IID interval, $[0.00299, 0.03074]$, is positive; its family-cluster interval, $[-0.00519, 0.03932]$, contains zero (Table 3). Thus the originally specified H2 existence rule is satisfied descriptively; it does not control family-wise error over ten intervals. Both diagnostic seeds reproduce the two labels and A5 stores their endpoints.

Nine of ten fixed comparisons retain their label: three other DynaMath pairs, all three MMBench pairs, and all three VB pairs. Every pair label is stable under both diagnostic seeds. The null proportion matters: these data show that a decision *can* change, not that model-comparison conclusions usually change. We report all ten intervals in Appendix B.

Because H2 is a pointwise descriptive rule over ten intervals, we added a post-hoc sensitivity using nominal 99.5% Bonferroni-adjusted percentile intervals and 100,000 replicates per interval. The ad-

Table 3: The changed pair depends on multiplicity contract. The adjusted row uses nominal 99.5% per-pair percentile-bootstrap intervals. Bonferroni gives at least 95% family-wise coverage only if each marginal interval attains its nominal coverage; this audit does not test that premise.

Pair / nominal level	Gap	Cell-IID interval	Family interval	Labels
Dyna: InternVL2-26B − 8B Pointwise 95%	0.01677	$[0.00299, 0.03074]$	$[-0.00519, 0.03932]$	positive / unresolved
MMB: LLaVA1.5-13B − 7B Bonferroni 99.5%	0.01442	$[0.00428, 0.02465]$	$[-0.00177, 0.03078]$	positive / unresolved

justment is a multiplicity contract, not an empirical coverage validation. The number of cell/family label differences remains one, but the identity changes: the DynaMath pair becomes unresolved under both contracts, while MMBench LLaVA1.5-13B minus 7B is positive under cell-IID and unresolved under family resampling (Table 3). This pattern is stable under two additional 100,000-replicate seeds. We therefore treat the original DynaMath change as a pointwise illustration, not a multiplicity-robust discovery.

4.3 Robustness and negative evidence

All 23 primary ratio-band labels remain unchanged under the two diagnostic seeds. DynaMath leave-one-position-out ratios remain above two for every model and omitted position (overall 2.051–2.370), ruling out a single variant column as the driver. MMBench's all-3,217-family unequal-size sensitivity yields ratios of 1.728–1.874, consistent with the constant-four primary subset. Corrected VB complete-three-cell sensitivities yield 0.853, 0.920, and 0.872: direction remains non-widening, although one secondary equivalence-band label is seed-unstable and is not promoted to a claim. No MMBench or VB fixed pair changes label under the primary pointwise 95% comparison. This is not a claim of stability to changing the included samples. Exploratory checks added during AI-assisted review change both labels to positive for MMBench InternVL2-26B minus 8B on all families and VB Gemma3 minus Qwen3 on their 59 jointly complete families. Appendix F reports all six tested MMBench/VB subset comparisons, including unchanged outcomes.

A post-hoc analytic sensitivity reaches the same precision pattern without percentile quantiles. Median family/cell design effects are 5.446 for DynaMath, 3.360 for MMBench, and 0.766 for VB; their square-root analytic SE ratios are 2.334, 1.833, and 0.875, closely tracking the bootstrap width-ratio medians. For VB, this is a variable-size pooled-statistic delta method conditional on observed eligible cells and family sizes; it does not correct selection from missing cells. Appendix E gives the formulas and per-model values; we do not assert a one-way ICC or a causal negative-dependence mechanism. A further post-hoc control fixes the number of cells in each variant column on complete supports: all 23 model matrices and ten paired-difference matrices retain the direction of the family-versus-pooled conditional SE comparison (Appendix F.2). These are correlated descriptive checks of conditional moments, not additional primary tests or validated population intervals.

5 Discussion

Scoring and resampling units. Scoring and resampling units serve different roles. A benchmark can score each perturbation, rotation, or counterfactual while keeping cells from the same source together when resampling. The resampling unit should reflect the target population and the assumed sampling structure. Our analysis of the released matrices shows that this choice can either widen or narrow the interval around an unchanged observed score. VB's complete-support composition check separates variation in column counts from empirical cross-column covariance: the InternVL3 family/stratified SE ratio is close to one, whereas Gemma3 and Qwen3 remain below one (Appendix F.2). This is a descriptive decomposition, not a causal explanation of the primary available-cell result.

Recent rank-uncertainty work links benchmark populations and inferential units to different evaluation questions Neuhof and Benjamini (2026a;b). Our audit measures the size and direction of the

fixed-statistic interval change across three source-derived VLM matrix formats. Reporting all ten model pairs shows both the changed comparison and the nine unchanged comparisons. The post-hoc Bonferroni analysis further shows why a pointwise example and a multiplicity-adjusted example must be identified separately.

A minimal reporting contract. For repeated-variant benchmarks, we recommend reporting (i) the source-family key and number of cells per family; (ii) the population to which uncertainty is intended to generalize; (iii) the observed scoring statistic, its intended population parameter, and whether it is cell- or family-weighted; (iv) an interval that resamples the declared unit; and (v) a fixed-statistic sensitivity under at least one plausible alternative unit. If model comparisons are important, pairs should be fixed independently of observed gaps, all null comparisons retained, and pointwise versus multiplicity-adjusted nominal levels stated without implying unvalidated coverage. These additions do not require retraining or repeated model calls when response matrices are released.

Interpreting repeated-variant evidence. Controlled variants expose how performance changes around a common source. Additional cells contribute to the observed cell-weighted score, but do not themselves establish independent replication; preserving families during resampling need not widen an interval. Reporting the source-family definition alongside the fixed-statistic sensitivity makes these assumptions visible when benchmark results inform model selection.

6 Limitations and Ethics

This audit covers three version-pinned public snapshots, 23 model artifacts, and ten fixed pairs; it is not a prevalence estimate over all VLM benchmarks. MMBench is intentionally a rotated-prompt cell-accuracy diagnostic and not its official CircularEval headline. Only three VB model outputs were public at the pinned commit; A5 corrects A4's erroneous omission of three noncanonical but officially retained labels. More importantly, editable A1–A4 timestamp fields conflict with local run timestamps and lack a trusted external anchor, so the preserved local audit package cannot independently establish prospective or confirmatory status. We accordingly present A5 as a corrected descriptive re-analysis while preserving the full contradictory history. Percentile bootstrap intervals describe source-corpus resampling variation, not prompt-generation, judge, training, API, or repeated-run uncertainty; without known population truth, we do not claim coverage optimality. The analytic and Bonferroni checks are post-hoc robustness evidence. The 10% width band is an original protocol convention rather than a universal standard. The paper also does not construct rank confidence intervals or propagate uncertainty across tasks as recent leaderboard work does Neuhof and Benjamini (2026b;a). Finally, source families can be crossed with other dependencies (templates, annotators, images, or episodes), for which a one-way family bootstrap may be incomplete; "family" is a tested unit, not asserted as the final independent unit.

The study uses only public benchmark artifacts, makes no model/API calls, and contains no human subjects or private data. DynaMath declares MIT terms in its pinned README, VB includes an MIT license, and OpenVLMRecords declares Apache-2.0; the local audit package retains this provenance. Its risk is primarily interpretive: a narrow audit could be used to criticize benchmark authors or to claim universal invalidity. We avoid that framing, preserve opposite-direction and null results, credit upstream artifacts, and distinguish published metrics from our diagnostics. Prepared metadata contain hashes and family keys but no credentials or additional personal information.

7 Conclusion

Across 23 repeated-variant VLM response matrices from three benchmarks, changing only the resampling unit materially changes interval width in benchmark-specific directions. Family intervals widen substantially for DynaMath and an MMBench cell diagnostic and narrow for VB. One of ten pointwise labels changes; a Bonferroni sensitivity retains one change but for a different pair. These results support a modest standard: benchmarks should declare the target population and tested resampling unit, then show whether uncertainty and important comparisons survive a plausible family-level analysis.

AI USE STATEMENT

Generative AI tools (OpenAI Codex, ChatGPT, and DeepSeek) assisted research planning, literature review, code development and execution, statistical analysis, writing, formatting, and reviewer-style checks. Scores derive from pinned public benchmark outputs; this audit generated no new model responses. Sources and calculations were checked against primary materials and reproducible artifacts. Review-added analyses are labeled exploratory. The authors are responsible for the accuracy, attribution, and conclusions of the work.

REPRODUCIBILITY STATEMENT

Appendix A reports every model matrix and Appendix B reports all fixed pairs. The anonymous supplementary archive contains the pinned public inputs, A0–A5 history, scorer and provenance amendments, deterministic runners, complete result objects, exported tables, and a reproduction verifier. The fixed-composition extension, its complete results and support keys, and a verification script are included as a separately labeled exploratory addendum.

REFERENCES

Eytan Bakshy and Eitan Frachtenberg. Design and analysis of benchmarking experiments for distributed internet services. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

Haodong Chen, Qiang Huang, Jiaqi Zhao, Qiuping Jiang, Xiaojun Chang, and Jun Yu. Measuring social bias in vision-language models with face-only counterfactuals from real photos. *arXiv preprint arXiv:2601.06931v2*, 2026. Template bootstrap: Appendix C.3.

J. de Curto, Dayani Plasencia, Diego Sanchez, and I. de Zarza. Visual grounding in zero-shot vision-language control. *arXiv preprint arXiv:2608.06154*, 2026.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, Yuxuan Qiao, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *ACM Multimedia*, 2024.

Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984.

Anik Jha. Whose gold? Annotator-pool disagreement is large at the item level, and hidden by small leaderboards. *arXiv preprint arXiv:2608.15980*, 2026.

Yifan Jiang, Dae Yon Hwang, Jesse C. Cresswell, and Freda Shi. Chartographer: Counterfactual chart generation for evaluating vision-language models. *arXiv preprint arXiv:2605.27311*, 2026.

Feixiang Liu, Shiye Wang, Qiang Qiu, and Zheng Wang. Witness evidence portfolios: Single-prefill risk detection for closed multimodal answers. *arXiv preprint arXiv:2607.27667*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint arXiv:2501.04234*, 2025.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. In *ICML Workshop on Hypothesis Testing*, 2026a. arXiv:2607.16259.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. *arXiv preprint arXiv:2606.08679*, 2026b.

Neil Tripathi. VB: Visibility benchmark for visibility and perspective reasoning in images. *arXiv preprint arXiv:2603.06680*, 2026.

Qiaoyuan Zheng and Yiqu Yang. Are near-tied LLM rankings robust to family-DIF-guided benchmark recomposition? *arXiv preprint arXiv:2609.00482*, 2026.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025.

A Complete Model-Level Results

All 23 primary matrices are reported below. G/N gives families/eligible cells. Intervals are pointwise 95%; widths and full precision are also exported in the supplement. Model abbreviations IVL, Q2VL, and LLaVA refer to the same exact model IDs in the released CSVs; the supplement retains full names.

Table 4: Every primary model: observed score, both intervals, and width ratio.

Benchmark	Model	G/N	Score	Cell-IID CI	Family CI	R
Dyna	GPT4o	501/5010	0.6333	[0.6202, 0.6467]	[0.6016, 0.6653]	2.3987
Dyna	Gemini1.5-Pro	501/5010	0.6040	[0.5906, 0.6176]	[0.5717, 0.6361]	2.3928
Dyna	IVL2-26b	501/5010	0.4102	[0.3968, 0.4240]	[0.3788, 0.4419]	2.3231
Dyna	IVL2-40B	501/5010	0.4186	[0.4050, 0.4321]	[0.3866, 0.4509]	2.3676
Dyna	IVL2-76B	501/5010	0.5361	[0.5224, 0.5497]	[0.5030, 0.5695]	2.4307
Dyna	IVL2-8b	501/5010	0.3934	[0.3800, 0.4070]	[0.3623, 0.4248]	2.3185
Dyna	Qwen2-72B	501/5010	0.5511	[0.5375, 0.5649]	[0.5176, 0.5848]	2.4599
Dyna	Qwen2-7B	501/5010	0.4214	[0.4076, 0.4351]	[0.3898, 0.4537]	2.3188
Dyna	claude3.5	501/5010	0.6483	[0.6353, 0.6613]	[0.6156, 0.6806]	2.5077
Dyna	deepseek7B	501/5010	0.2162	[0.2048, 0.2275]	[0.1906, 0.2433]	2.3158
Dyna	llama-3.2-90B	501/5010	0.4429	[0.4291, 0.4565]	[0.4112, 0.4750]	2.3353
Dyna	llava1.5-7B	501/5010	0.1639	[0.1535, 0.1741]	[0.1411, 0.1866]	2.2136
Dyna	LLaVA1.6-34B	501/5010	0.2719	[0.2597, 0.2842]	[0.2449, 0.2996]	2.2278
Dyna	LLaVA1.6-V-13B	501/5010	0.1964	[0.1852, 0.2074]	[0.1725, 0.2212]	2.1982
MMB	IVL2-26B	2688/10752	0.8564	[0.8496, 0.8630]	[0.8442, 0.8685]	1.8125
MMB	IVL2-8B	2688/10752	0.8510	[0.8442, 0.8578]	[0.8386, 0.8632]	1.8082
MMB	Q2VL-2B	2688/10752	0.8055	[0.7980, 0.8129]	[0.7921, 0.8190]	1.8062
MMB	Q2VL-7B	2688/10752	0.8564	[0.8498, 0.8631]	[0.8440, 0.8686]	1.8462
MMB	LLaVA1.5-13b	2688/10752	0.7528	[0.7447, 0.7608]	[0.7377, 0.7679]	1.8730
MMB	LLaVA1.5-7b	2688/10752	0.7384	[0.7301, 0.7466]	[0.7231, 0.7534]	1.8418
VB	Gemma3-12b	100/300	0.5430	[0.4925, 0.5937]	[0.4997, 0.5872]	0.8649
VB	IVL3-8b	100/300	0.4977	[0.4465, 0.5495]	[0.4490, 0.5465]	0.9466
VB	Qwen3-8b	100/251	0.5093	[0.4525, 0.5655]	[0.4626, 0.5581]	0.8456

B All Fixed Pair Results

Pair IDs below identify the same ordered model differences in both tables. G/N denotes shared families/cells. P, N, and U mean positive, negative, and unresolved. VB pairs use pair-specific intersections: gaps need not equal differences of the marginal scores above, and the three gaps do not define a transitive ranking on one common support.

Table 5: Exact pair mapping (left minus right); full released identifiers are preserved in the CSV.

ID	Benchmark	Left	Right
D1	Dyna	IVL2-26b	IVL2-8b
D2	Dyna	IVL2-76B	IVL2-40B
D3	Dyna	Qwen2-72B	Qwen2-7B
D4	Dyna	LLaVA1.6-34B	LLaVA1.6-V-13B
M1	MMB	IVL2-26B	IVL2-8B
M2	MMB	LLaVA1.5-13b	LLaVA1.5-7b
M3	MMB	Q2VL-7B	Q2VL-2B
V1	VB	Gemma3-12b	IVL3-8b
V2	VB	Gemma3-12b	Qwen3-8b
V3	VB	IVL3-8b	Qwen3-8b

Table 6: All ten primary pointwise 95% paired intervals, with actual shared sample sizes.

ID	G/N	Gap	Cell-IID CI	Family CI	Labels
D1	501/5010	0.0168	$[0.0030, 0.0307]$	$[-0.0052, 0.0393]$	P/U
D2	501/5010	0.1176	$[0.1038, 0.1317]$	$[0.0920, 0.1437]$	P/P
D3	501/5010	0.1297	$[0.1158, 0.1437]$	$[0.1048, 0.1551]$	P/P
D4	501/5010	0.0754	$[0.0621, 0.0892]$	$[0.0513, 0.1000]$	P/P
M1	2688/10752	0.0054	$[-0.0006, 0.0115]$	$[-0.0043, 0.0153]$	U/U
M2	2688/10752	0.0144	$[0.0073, 0.0215]$	$[0.0032, 0.0259]$	P/P
M3	2688/10752	0.0509	$[0.0442, 0.0577]$	$[0.0403, 0.0616]$	P/P
V1	100/300	0.0453	$[-0.0173, 0.1092]$	$[-0.0115, 0.1018]$	U/U
V2	100/251	0.0357	$[-0.0277, 0.1002]$	$[-0.0148, 0.0867]$	U/U
V3	100/251	-0.0178	$[-0.0816, 0.0454]$	$[-0.0683, 0.0311]$	U/U

Table 7: All ten nominal 99.5% Bonferroni intervals (100,000 replicates, seed 20260810). Stability refers to labels under the two additional 100,000-replicate seeds. Nominal family-wise coverage remains conditional on marginal coverage, which is not validated here.

ID	Cell-IID CI	Family CI	Labels	Seed-stable
D1	$[-0.0030, 0.0367]$	$[-0.0152, 0.0489]$	U/U	true
D2	$[0.0976, 0.1377]$	$[0.0812, 0.1557]$	P/P	true
D3	$[0.1098, 0.1499]$	$[0.0942, 0.1659]$	P/P	true
D4	$[0.0561, 0.0948]$	$[0.0407, 0.1108]$	P/P	true
M1	$[-0.0032, 0.0140]$	$[-0.0086, 0.0193]$	U/U	true
M2	$[0.0043, 0.0246]$	$[-0.0018, 0.0308]$	P/U	true
M3	$[0.0414, 0.0605]$	$[0.0359, 0.0660]$	P/P	true
V1	$[-0.0452, 0.1362]$	$[-0.0365, 0.1260]$	U/U	true
V2	$[-0.0561, 0.1269]$	$[-0.0365, 0.1095]$	U/U	true
V3	$[-0.1079, 0.0725]$	$[-0.0896, 0.0523]$	U/U	true

The supplementary `pair-seed-intervals.csv` and `model-seed-diagnostics.csv` export all recorded primary-analysis diagnostic endpoints. These are distinct from secondary VB diagnostics, which recorded ratios only.

C Eligibility and Scorer Denominators

All six MMBench files have 3,217 families and 12,175 rows. The common four-rotation primary subset has 2,688 families and 10,752 rows per model. The remaining 529 families comprise 164 two-cell and 365 three-cell families, totaling 1,423 rows per model. Both stages use the same pinned matcher; unresolved outputs are incorrect, with no silently dropped rows.

Table 8: MMBench fallback counts. Denominators in the headings are per model.

Model	Primary / 10,752	All / 12,175
IVL2-26B	0	0
IVL2-8B	0	0
LLaVA1.5-13b	0	0
LLaVA1.5-7b	0	0
Q2VL-2B	170	200
Q2VL-7B	101	102

All primary VB models represent 100 families. Gemma3 and InternVL3 have 300 eligible headline cells each; Qwen3 has 251. Qwen3 is missing 1 BASE, 36 TEXT_FLIP, and 12 IMAGE_FLIP cells; DOUBLE_FLIP is missing another 17 cells. These are parse/scoring exclusions, not assumed random losses. The supplementary `support-keys/` files identify the exact per-model and per-pair keys and their hashes. No inference about accuracy on omitted cells is made.

D All Declared VB Secondary Results

Complete-three uses each model's own fully eligible families. Available-four adds the diagnostic DOUBLE_FLIP relation and keeps every eligible cell; Qwen3 is not restricted to complete four-cell families. Available-four changes the evaluated support for all three models; complete-three changes it only for Qwen3. These secondary results do not replace the primary headline-cell analysis.

Table 9: VB secondary scores and all recorded ratio seeds. All three base seeds (20260810, 20260911, and 20261012) use 5,000 replicates for these secondary results. Base-seed (20260810) interval endpoints are in the CSV; diagnostic-secondary endpoints were not stored and are not invented here.

Support	Model	G/N	Score	R_0	R_1	R_2
Complete-three	Gemma3-12b	100/300	0.5430	0.8532	0.8721	0.9057
Complete-three	IVL3-8b	100/300	0.4977	0.9203	0.9237	0.9234
Complete-three	Qwen3-8b	59/177	0.4555	0.8725	0.7997	0.8512
Available-four	Gemma3-12b	100/400	0.5500	0.8348	0.8190	0.8129
Available-four	IVL3-8b	100/400	0.5749	0.8663	0.8833	0.8882
Available-four	Qwen3-8b	100/334	0.5750	0.9923	1.0077	1.0351

For Gemma3 and InternVL3, complete-three retains exactly the primary keys, scores, and ordering; these rows are not additional evidence of robustness to support changes. Their ratios come from separate Monte Carlo computations: the primary run uses 20,000 replicates, whereas all secondary runs use 5,000. Analysis-specific labels (`VB` versus `VB_complete_three`) derive different random seeds. The primary family sampler draws family indices and pools sums and sizes; complete-three draws multinomial counts over family means. For these equal-sized families the target empirical bootstrap distributions coincide, but the sampled sequences differ. A deterministic check with four shared index arrays per model found the two statistics equal to floating-point precision; it computed no new intervals. Details are in `reporting-corrections-20260916/`.

Gemma3 complete-three crosses the 0.90 equivalence-band boundary at the second diagnostic seed. Qwen3 available-four stays within the equivalence band, but crosses 1: its direction is not uniformly narrowing across seeds. Band stability must not be read as direction stability.

E Existing Analytic Variance Sensitivity

Let $N = \sum_g m_g$, $T_g = \sum_j y_{gj}$, and let s_y^2 denote the sample variance (ddof=1) of eligible cells. The cell variance estimate is $V_I = s_y^2/N$. For constant family size, $V_F = s_{\bar{y}_g}^2/G$. For unequal family sizes the existing post-hoc delta-method check uses

$$V_F = \frac{G}{G-1}\frac{\sum_g (T_g - \widehat{\mu} m_g)^2}{N^2}, \qquad \mathrm{DE} = V_F/V_I. \tag{3}$$

This is conditional on the observed eligible-cell structure, not a missing-data correction. The formulas and all analytic values below come from a post-hoc calculation on corrected A5 scores and primary supports, recorded in this supplement file:

`experiments/VSUA-POSTHOC-SENSITIVITY-A1.json`

Its SHA-256 prefix is `64993918f854`. This post-hoc revision suffix A1 is distinct from the stopped original run `VSUA-FORMAL-A1`; it does not identify that earlier run. Absolute bootstrap widths use percentile endpoints and are not constrained to a normal-quantile multiple.

Table 10: Absolute 95% percentile widths and existing analytic design effects for every primary model.

Benchmark	Model	IID width	Family width	DE	$\sqrt{DE}$
Dyna	GPT4o	0.0265	0.0637	5.8151	2.4115
Dyna	Gemini1.5-Pro	0.0269	0.0645	5.6022	2.3669
Dyna	IVL2-26b	0.0272	0.0631	5.3837	2.3203
Dyna	IVL2-40B	0.0271	0.0643	5.5195	2.3494
Dyna	IVL2-76B	0.0273	0.0665	5.8724	2.4233
Dyna	IVL2-8b	0.0269	0.0625	5.4019	2.3242
Dyna	Qwen2-72B	0.0273	0.0673	5.9370	2.4366
Dyna	Qwen2-7B	0.0275	0.0639	5.4901	2.3431
Dyna	claude3.5	0.0259	0.0651	5.9628	2.4419
Dyna	deepseek7B	0.0228	0.0527	5.2963	2.3014
Dyna	llama-3.2-90B	0.0274	0.0639	5.3898	2.3216
Dyna	llava1.5-7B	0.0206	0.0455	4.9321	2.2208
Dyna	LLaVA1.6-34B	0.0246	0.0547	4.9188	2.2178
Dyna	LLaVA1.6-V-13B	0.0222	0.0487	5.0162	2.2397
MMB	IVL2-26B	0.0134	0.0243	3.3732	1.8366
MMB	IVL2-8B	0.0136	0.0246	3.3466	1.8294
MMB	Q2VL-2B	0.0149	0.0269	3.2784	1.8106
MMB	Q2VL-7B	0.0133	0.0246	3.3944	1.8424
MMB	LLaVA1.5-13b	0.0161	0.0301	3.3742	1.8369
MMB	LLaVA1.5-7b	0.0165	0.0303	3.3303	1.8249
VB	Gemma3-12b	0.1012	0.0875	0.7660	0.8752
VB	IVL3-8b	0.1030	0.0975	0.8836	0.9400
VB	Qwen3-8b	0.1130	0.0955	0.7312	0.8551

F Exploratory Analyses Added During Review

The following results were computed during AI-assisted review on September 16, after the primary results were known. They were not part of the original frozen analyses. The scripts and outputs are supplied in a separate review addendum. No claim of prospective selection, validated coverage, or causal mechanism follows from these checks.

F.1 Changing the included samples

We report all six reviewed subset comparisons: all families for the three MMBench pairs, and jointly complete three-cell families for the three VB pairs. No DynaMath pair subset check is added here. Each row holds its support fixed between resampling schemes. Relative to the primary analysis, support changes for M1–M3 and V2–V3; V1 retains its original support. These exploratory rows do not revise the primary 1/10 count.

Table 11: Every reviewed subset comparison: 20,000 replicates, seed 20260916, pointwise 95%. IDs match Appendix B.

ID	G/N	Gap	Cell-IID CI	Family CI	Labels
M1	3217/12175	0.0120	$[0.0063, 0.0177]$	$[0.0030, 0.0212]$	P/P
M2	3217/12175	0.0133	$[0.0067, 0.0202]$	$[0.0026, 0.0239]$	P/P
M3	3217/12175	0.0581	$[0.0516, 0.0645]$	$[0.0480, 0.0684]$	P/P
V1	100/300	0.0453	$[-0.0185, 0.1095]$	$[-0.0128, 0.1022]$	U/U
V2	59/177	0.0823	$[0.0130, 0.1525]$	$[0.0232, 0.1441]$	P/P
V3	59/177	-0.0089	$[-0.0884, 0.0695]$	$[-0.0664, 0.0456]$	U/U

M1 and V2 change from U/U on the primary samples to P/P on the new samples. All other reviewed pairs keep their labels. V1 still uses 100 families, whereas V2 and V3 use 59: this is not a three-model common-support analysis. Two additional 5,000-replicate seeds preserve the VB subset labels, but their endpoints were not stored. MMBench subset intervals have only one reviewed seed; multi-seed stability is untested.

F.2 A FIXED-COLUMN COMPOSITION CHECK

We extend the post-hoc check to all 23 model matrices and the ten fixed pairs on complete supports. DynaMath and MMBench retain their primary supports. VB uses each model's complete-three subset (100, 100, and 59 families); its three pairs use pair-specific intersections of 100, 59, and 59 families. The two Qwen3 pairs each exclude 41 families and 74 previously available cells (251 to 177). These subsets do not replace the primary available-cell comparisons or their decision-change count.

For a complete $G \times m$ matrix y, consider pooled-cell resampling of Gm cells, independent resampling of G cells within each column, and resampling of G complete families. Each scheme has conditional expectation equal to the observed cell mean. Their exact empirical bootstrap variances, with Var_0 and Cov_0 denoting ddof=0 moments, are

$$V_p = \mathrm{Var}_0(y)/(Gm), \qquad V_s = \sum_j \mathrm{Var}_0(y_{\cdot j})/(Gm^2), \qquad V_f = \mathrm{Var}_0(\bar{y}_g)/G. \tag{4}$$

Writing $K = \mathrm{Var}_0(\bar{y}_{\cdot j})/(Gm)$ and $C = 2\sum_{j<k} \mathrm{Cov}_0(y_{\cdot j}, y_{\cdot k})/(Gm^2)$, standard variance expansion gives $V_p - V_s = K \geq 0$, $V_f - V_s = C$, and $V_f - V_p = C - K$. Thus a smaller family variance than pooled variance need not imply negative column-centered covariance: a larger composition term can outweigh positive covariance. These are classical identities, not a new bootstrap method; preserving balance in benchmark resampling has established precedents Bakshy and Frachtenberg (2015). We evaluate the analytic expressions in float64; no bootstrap simulation is required.

Table 12: Post-hoc medians of conditional bootstrap SD ratios. These are not percentile-interval width ratios. Model and pair sets are correlated descriptive views, not independent repetitions.

Benchmark	Matrix	n	$\sqrt{V_f/V_p}$	$\sqrt{V_f/V_s}$
DynaMath	model	14	2.3315	2.3321
MMBench	model	6	1.8328	1.8328
VB	model	3	0.8723	0.8942
DynaMath	pair	4	1.7864	1.7873
MMBench	pair	3	1.5951	1.5951
VB	pair	3	0.8704	0.9116

All 20 DynaMath/MMBench model matrices retain $V_f > V_s$; all three VB model matrices retain $V_f < V_s$. For VB, the family/stratified SD ratios are 0.8743 (Gemma3), 0.9925 (InternVL3), and 0.8942 (Qwen3). InternVL3 is close to one: $K/V_p = 0.10910$ and $C/V_p = -0.01326$ on its complete subset. The other two have $C/V_p = -0.23460, -0.17460$ and $K/V_p = 0.00451, 0.12845$, respectively. None of the 23 model matrices or ten paired-difference matrices reverses the direction of its pooled comparison when using the stratified control. This count describes only the audited, correlated matrices.

The control fixes column counts while breaking family alignment across columns; it does not establish a preferred population sampling scheme. DynaMath columns denote released variant slots rather than verified semantic roles. Full column moments, covariance matrices, support keys, and per-pair results are supplied in `composition-20260916/` in the anonymous supplement. These conditional moments neither validate coverage nor isolate a causal explanation, and they do not transfer directly to unequal-size available-cell ratios. The paper's percentile-width tolerance band is not applied to these SD ratios.

G REPRODUCIBILITY AND AMENDMENT LEDGER

The original A5 runner uses NumPy PCG64 with base seed 20260810, 20,000 primary replicates, linear quantile interpolation, and paired differences. Its primary diagnostic base seeds are 20260911 and 20261012 (5,000 replicates). Each CI call initializes a separate generator from the first four SHA-256 bytes of `base:label`, interpreted as an unsigned big-endian integer; labels distinguish analysis, model or pair, and resampling scheme. There is no advancing global RNG stream. All VB secondary runs, including base seed 20260810, use 5,000 replicates. A5 result, trace, and upstream

manifest SHA-256 prefixes remain `9750a3d186ae`, `1be7534a75e1`, and `bea3086fffba`; full hashes are in the supplement.

A0–A3 stopped before effect estimation: two DynaMath categorical-token issues, MMBench free-form predictions, and VB complete-three availability below the initial threshold. A separate AI-assisted review of A4 identified the noncanonical-label omission and conflicting human-readable timestamps. A5 corrected our own earlier scorer after those outcomes were known. A0–A5 artifacts remain unchanged. The September 16 review outputs are separate, explicitly exploratory additions; they do not restore prospective status.

The review addendum contains per-model and pair intervals, recorded seed diagnostics, secondary results, exact support keys, and the subset and fixed-relation scripts. The additional `composition-20260916/` directory extends the conditional-moment control to all 23 models and ten fixed pairs; it was computed on September 16 and integrated into this manuscript on September 21. Its standalone verifier reconstructs those moments from the packaged public inputs without rerunning the primary bootstrap. It does not supply endpoints that earlier scripts did not save. AI reviews and local reproducibility checks do not establish original upstream execution fidelity or population coverage.