BALANCE OF BENCHMARKS: EVIDENCE RETENTION AND MULTIPLICITY-AWARE MODEL COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark aggregation couples two quantities that should be set separately: how much evidence a comparison uses and how much influence each benchmark receives. Curated suites control influence by selecting benchmarks, potentially excluding informative observations, while equal weighting of a broader pool lets repeated or closely related benchmarks gain influence through sheer numbers. We introduce Balance of Benchmarks (BoB), which retains eligible observations, fits benchmark-specific response curves to place different difficulty ranges on a shared ability scale, and down-weights benchmarks in crowded semantic neighborhoods using description similarity. We evaluate evidence retention and density weighting through separate interventions on WildScores, a sparse collection of 148 developer-reported benchmarks, with 14 Artificial Analysis benchmarks as a controlled reference. Holding fitting models fixed and withholding source-lineage families, expanding one historical ten-benchmark suite to the full eligible pool raises BoB's held-out Spearman correlation by 0.035 over 50 targets (Holm $p = 0.013$). Density weighting reduces sensitivity to complete re-listings: after four copies are added, mean absolute rank displacement falls from 1.95 under equal-weight equating to 0.63. Reporting masks reveal a further distinction: conserving a group's column weight leaves each model's observed weight dependent on its reporting coverage. Together, these results clarify how evidence retention, benchmark multiplicity, and reporting coverage jointly shape model comparison.

1 INTRODUCTION

Which benchmark observations should a model comparison use, and how much weight should each receive? These are different design choices. Curated suites control capability emphasis by selecting evaluations and assigning category weights, as illustrated by the Artificial Analysis (AA) Intelligence Index (Artificial Analysis, 2026b). Selection can nevertheless exclude informative observations. A benchmark that saturates on leading models may still distinguish models near the floor of a harder evaluation. Our question is whether an aggregation procedure can use such additional evidence while explicitly moderating the influence of repeated or closely related benchmarks.

We study aggregation of already reported benchmark-level scores. Public model-release reports supply a sparse matrix whose columns differ in difficulty, capability coverage, reporting frequency, and evaluation protocol. The available observations are aggregate scores with benchmark descriptions and provenance. This measurement resolution makes reporting coverage and score comparability part of the aggregation problem. The predictive target is performance on held-out reported benchmarks among eligible release groups.

The distinction becomes concrete when a benchmark is listed twice. Equal weighting gives its scores extra weight even though the second listing adds no measurement. Related benchmarks can instead provide fresh measurements, making the choice of capability emphasis distinct from the amount of evidence collected. Sparse reporting adds a further complication. Sharing a fixed total weight across related columns gives different observed group weights to models evaluated on different subsets. Benchmark weights, model-specific observed weights, and changes in fitted rankings are therefore three distinct objects to analyze.

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.

Balance of Benchmarks (BoB) combines three components for this setting. An eligibility screen admits scores with documented provenance and a normalizable scale. Score equating fits one response curve per benchmark so that easy and hard benchmarks share an ability scale. Inverse-density weighting assigns smaller coefficients to benchmarks whose descriptions sit in crowded semantic neighborhoods. Benchmark-level response models and clone-robust weighting have established precedents (Ho et al., 2025; Berriaud & Wattenhofer, 2026a;b). Our focus is how these components interact under sparse reporting. For a fixed eligible benchmark list, the weights are computed from descriptions alone. Their effect on the fitted ability θ_m also depends on the fitted curves and on which models report which benchmarks. An optional task-conditioned extension is described and evaluated separately in Appendix F.

We test the two choices with separate interventions on WildScores, using AA as a denser reference. For retention, we expand a curated suite of ten benchmarks from the historical AA Index v3.0 list (Artificial Analysis, 2026b) to the full eligible pool while keeping the model population fixed and withholding entire source-lineage families. Held-out ranking improves under both BoB and equal-weight equating. For influence control, we re-list existing benchmarks with their complete scores and observation masks, adding no measurements. Density weighting reduces mean rank displacement by 67.9% relative to equal-weight equating, and the effect carries over to an affine-factor estimator. We then analyze the reporting and fitting conditions under which this weighting translates into stable model comparisons.

Contributions. **(i) An integrated formulation for sparse reported scores.** BoB makes evidence eligibility, difficulty equating, and semantic weighting explicit and reproducible within one aggregation procedure. **(ii) Separate tests of retention and influence control.** Fixed-population held-out prediction and exact or paraphrased re-listings distinguish evidence-pool gains from sensitivity to benchmark multiplicity. Comparisons include observed-cell factors and same-information semantic grouping, keeping estimator choice separate from the retention contrast. **(iii) Analysis of reporting and fitting effects.** Conserving a group's column weight can change each model's observed weight. We derive this reporting-mask effect and trace residual ranking changes to reporting masks, neighboring weights, initialization, and refitting. Metadata restriction and estimator transfer establish the empirical scope of the resulting comparisons.

2 Related work

Benchmark selection and evidence efficiency. Holistic suites organize broad evaluations (Liang et al., 2023; Srivastava et al., 2023), while AA combines selected benchmarks with prescribed weights (Artificial Analysis, 2026a). Evaluation-compression methods select fewer measurements to preserve model comparisons (Perlitz et al., 2024a; Maia Polo et al., 2024; Kipnis et al., 2025; Vivek et al., 2024), and marginal-coverage approaches use prompt representations or score covariance to guide subset choice (Yao et al., 2026; Smola, 2026). These methods select evidence for efficient evaluation. BoB studies already reported scores, testing evidence retention on fixed model populations and assigning benchmark weights from descriptions.

Aggregation and benchmark influence. Arena aggregates human preferences (Chiang et al., 2024), with participation and selective disclosure affecting its evidence (Singh et al., 2025). Rank aggregation offers another way to combine benchmark results (Colombo et al., 2022), domain-specific taxonomies define explicit capability groups (Hudson, 2026), and score-matrix methods study agreement, dimensionality, and predictability (Perlitz et al., 2024b; Ruan et al., 2024; Sha & Zhao, 2026). Zhang & Hardt (2024) study ranking sensitivity and its trade-off with benchmark diversity. Directly addressing multiplicity, Berriaud & Wattenhofer (2026a) develop clone-robust task weights from metric similarity, with axioms governing symmetry, continuity, and locality. Their subsequent construction extends this framework to arbitrary metric spaces (Berriaud & Wattenhofer, 2026b). BoB uses a reciprocal kernel-density rule. Its effective mass is the order-zero similarity-sensitive diversity of Leinster & Cobbold (2012) at uniform frequencies, with related sums in the spread construction of Willerton (2015). We study this rule within sparse-score estimation, where neighboring descriptions, reporting masks, initialization, and refitting determine how column weights translate into ranking changes. Its weight-sharing property applies to isolated co-located entries (Section 3.2).

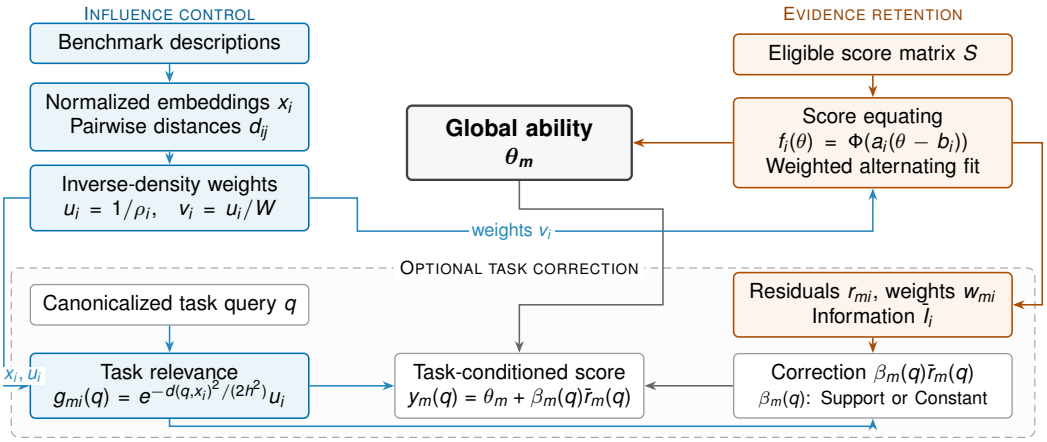

Figure 1: **What evidence enters and how much it counts.** Eligible scores enter score equating, while description-derived density weights set benchmark coefficients in the fitting objective. Scores, observation masks, curves, and robust cell weights jointly determine global ability θ_m. Held-out prediction tests retention, and re-listing tests ranking sensitivity. The dashed branch is an optional extension (Appendix F).

Score linking. Because benchmark scores differ in difficulty and saturation, characteristic curves place them on a common latent scale, following score-equating ideas from item response theory (Lalor et al., 2016; Rodriguez et al., 2021; Lord & Novick, 1968; Stocking & Lord, 1983; Kolen & Brennan, 2014). The Rosetta Stone framework underlying the Epoch Capabilities Index fits model capability, benchmark difficulty, and discrimination to partially observed aggregate scores, including externally reported evaluations (Ho et al., 2025). BoB builds on this benchmark-level modeling approach by combining score linking with description-based weighting. The contribution centers on their integration and empirical analysis: fixed-population retention comparisons and the consequences of multiplicity under sparse reporting.

3 METHOD

BoB implements the two design choices through three components (Figure 1). An eligibility screen selects which reported scores enter a partially observed matrix $S = (s_{mi})$, and score equating places them on a shared ability scale. Semantic inverse-density weighting assigns each benchmark a coefficient in the fitting objective, moderating the weight of crowded description neighborhoods. Together, they shape the fitted global ability θ_m. Qwen3-Embedding-8B (Zhang et al., 2025) maps descriptions to unit vectors x_i, compared by angular distance $d(\cdot, \cdot)$. Proximity is a proxy for related benchmark content, not a certificate of identical questions or evaluation protocols.

Eligible evidence. A score enters S when its source and evaluation setting are documented, its scale is normalized, and its benchmark and model rows meet minimum observation counts (Section B). A separate metadata restriction tests unresolved protocol differences within a column (Section 4.1).

3.1 SCORE EQUATING

Score equating models how benchmark scores vary across a shared latent ability scale. The characteristic curves represent saturation near zero and one, as illustrated by AIME and MATH-500 (Figure 2). We use a benchmark-level working model inspired by item response theory (Lord & Novick, 1968; Embretson & Reise, 2000), fitted to aggregate scores. Each benchmark has a monotone expected-score curve:

$$f_i(\theta) = \Phi\big(a_i(\theta - b_i)\big), \qquad r_{mi} = \frac{s_{mi} - f_i(\theta_m)}{\sigma_i}. \tag{1}$$

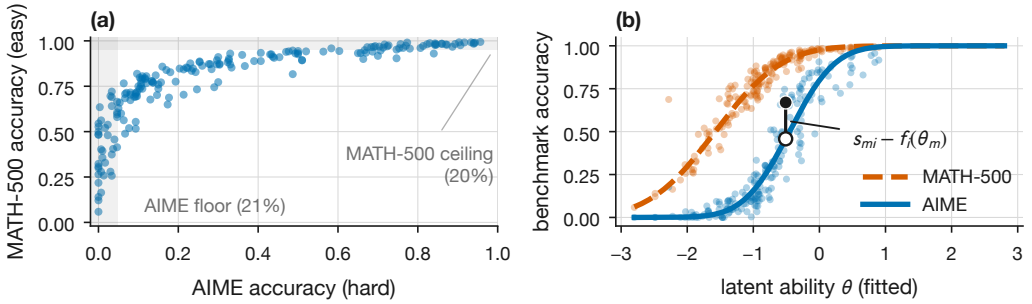

Figure 2: **Benchmark-specific curves place different difficulty ranges on a shared ability scale.** **(a)** AIME's floor and MATH-500's ceiling in the frozen AA snapshot. Shaded bands mark scores within 0.05 of the endpoints. **(b)** Curves give expected scores at shared ability. Filled and open markers show one model's observed and expected AIME scores, whose gap divided by σ_i gives r_{mi} (Equation (1)).

Here a_i and b_i describe discrimination and difficulty, Φ is the standard normal cumulative distribution function, and σ_i is the residual scale. A positive residual r_{mi} means above-expected performance at the model's fitted ability.

Alternating estimation fits the curves and updates ability by

$$\theta_m = \arg\min_{\theta} \sum_{i:\, s_{mi} \text{ observed}} w_{mi} v_i \big(s_{mi} - f_i(\theta)\big)^2, \tag{2}$$

where the robust weight w_{mi} limits the influence of outlying scores and v_i is the semantic coefficient of Section 3.2. After an ability update, we standardize jointly as $\theta' = (\theta - \mu)/s$, $a' = sa$ and $b' = (b - \mu)/s$, preserving every curve prediction. θ_m serves as the model's global level. Table 5 consolidates the full fitting procedure, with score-weight definitions in Section A.3.1. We stop after six rounds of curve fitting and ability updates. Extending fitting to 48 rounds lowers full-pool held-out Spearman in the fixed-population comparisons (Table 13), with stopping sensitivities reported in Section G.1.

3.2 Semantic inverse-density weighting

Under equal weighting, every listing of a benchmark receives a full coefficient. We weight each benchmark by the inverse of its local description density, so that entries in crowded neighborhoods share their weight:

$$k(d) = e^{-d^2/(2h_{\text{dens}}^2)}, \qquad \rho_i = \sum_j k(d_{ij}), \qquad u_i = \rho_i^{-1}, \qquad W = \sum_i u_i. \tag{3}$$

An entry retains more mass when its neighborhood is sparsely represented. Normalized weights are $v_i = u_i/W$, with the equivalent mean-one rescaling used in the fit. At fixed positive bandwidth and list size, these weights are strictly positive, permutation-invariant, and continuous in the distances. The construction makes benchmark multiplicity explicit:

Proposition 1 (Marginal mass at fixed semantic scale)**.** *Holding the kernel bandwidth fixed, adding a benchmark with similarities $k_i = k(d(x, x_i))$ changes the unnormalized effective mass by*

$$\Delta W = \frac{1}{1 + \sum_i k_i} - \sum_i \frac{k_i}{\rho_i(\rho_i + k_i)}.$$

Against one isolated benchmark at distance d, $\Delta W = (1 - k(d))/(1 + k(d))$.

In particular, isolated co-located entries share one unit of unnormalized mass. This guarantee concerns column weights only. Section 4.2 measures how far fitted rankings still move. We set $h_{\text{dens}} = 0.25 \operatorname{median}_{i<j} d_{ij}$ within each training pool and vary the multiplier in sensitivity analyses (Sections C.2 and G.1).

4 EXPERIMENTS

Each aggregation decision has a reference practice and its own intervention. **Q1, evidence retention:** with the model population and estimator fixed, does replacing a curated suite with all eligible columns improve held-out comparison? **Q2, influence control:** with measurements fixed, does density weighting reduce the ranking change that equal weighting gives repeated listings?

Throughout, BoB denotes the density-weighted, score-equated ability θ_m fitted with six rounds. All WildScores analyses condition on the scores developers chose to report. For held-out prediction on this corpus, Spearman correlation compares observed and predicted scores averaged within release groups. Correlations are then averaged over seeds within targets and over targets within benchmark families, with families weighted equally (Tables 1 and 2). We report every declared comparison, and Section 5 collects the resulting limitations.

Table 1: Frozen evaluation matrices. Configurations are separately listed model settings, and observed cells are model–benchmark scores.

Corpus	Configurations	Benchmarks	Observed cells	Observed (%)
Artificial Analysis	605	14	5,441	64.2
WildScores	354	148	5,371	10.3

Table 2: Evaluation map. Retention panels pair estimators on identical fitting rows within each panel. Re-listing scenarios measure sensitivity of fitted rankings to controlled changes in the recorded columns.

Analysis	Evaluation population	Intervention	Outcome
Q1, strict panel	50 targets, 29 families	Full pool versus curated suite, ≥ 3 curated scores/model	Family-mean held-out Spearman
Q1, broad panel	120 targets, 60 families	Same contrast, ≥ 1 curated score/model	Same ranking metric
Metadata screen	66 targets in the broad panel	Remove 38 columns with protocol conflicts	Same ranking metric
Q2, re-listing	354 configurations, eight source columns	Four complete or partial copies	Mean absolute rank displacement
Affine transfer	127 configurations (strict-panel rule)	Same copy interventions	Same displacement metric
AA reference	605 configurations, 14 targets	Leave one benchmark out	Level and residual profile

4.1 EVIDENCE RETENTION ON A FIXED MODEL POPULATION

Corpus and withholding. We compile WildScores, a dataset of 5,371 developer-reported scores covering 354 model configurations in 316 release groups and 148 benchmarks (10.3% coverage), with documented provenance and score normalization. Source review corrects documented scale and setting issues. Residual within-column protocol conflicts are tested by a separate metadata restriction below. For each of two fixed seeds, five outer blocks withhold entire source-lineage families, keeping benchmarks with shared question sources or derivations together. Family labels govern withholding and inference, with one explicitly metadata-assisted weighting control below. Section B details eligibility.

Fixed-population design. Fits to the curated suite and full pool use identical model rows within each panel, separating changes in the evidence pool from changes in the fitting population. The curated suite consists of ten benchmarks from the historical AA Index v3.0 list (Artificial Analysis, 2026b) and is distinct from the published AA composite. The contrast changes both the amount and domain composition of evidence, so it evaluates this pool expansion rather than pool size alone. Each fold withholds complete target source-lineage families. The strict panel requires at least three

Table 3: Evidence retention on fixed fitting populations. Entries are family-mean held-out Spearman correlations at six equating rounds. Panels differ in fitting population. Δ (Full − Curated) measures retention, while within-column comparisons assess complete estimators. Within the BoB group, $\rightarrow$ replaces density weighting. Paired effects and uncertainty are in Table 11. † uses source-lineage metadata unavailable to the description-only weighting rule. ‡ uses description-based complete linkage, with cluster count matched to rounded soft effective mass (Section E.1).

	Strict panel ($\geq$ 3 curated/model)			Broad panel ($\geq$ 1 curated/model)		
Method	Curated	Full	Δ	Curated	Full	Δ
Existing aggregation rules						
Uniform standardized averaging	0.819	0.827	+0.008	0.742	0.765	+0.023
Mean-imputed PC1	0.772	0.752	-0.019	0.624	0.652	+0.028
Observed-cell rank-one factor	0.819	0.830	+0.012	0.721	0.779	+0.058
Observed-cell affine factor	0.823	0.849	+0.026	0.764	0.815	+0.051
BoB and ablations						
BoB	0.804	0.839	**+0.035**	0.765	0.802	**+0.036**
Equal-weight equating	0.804	0.840	+0.037	0.765	0.799	+0.034
density $\rightarrow$ semantic groups‡	0.804	0.829	+0.026	0.765	0.798	+0.033
Metadata-assisted control						
density $\rightarrow$ source-family weights†	0.798	0.852	+0.054	0.762	0.806	+0.044

remaining curated observations per model and yields 50 targets in 29 families. The broad panel requires at least one and yields 120 targets in 60 families. Both require three observations per column and training-only variance filters. Comparisons are paired within panels because the populations differ.

Additional evidence benefits equating. Table 3 compares the curated suite and full pool on identical fitting populations within each panel. On a fixed model population, the full pool improves BoB by +0.0350 family-mean Spearman in the strict panel (95% CI [0.0183, 0.0520], Holm $p = 0.0130$) and by +0.0364 in the broad panel (Holm $p = 0.0107$). Equal-weight equating also benefits, by +0.0367 and +0.0342, respectively (Table 11). These matched comparisons show the value of evidence beyond the tested curated suite with equating. Density weighting's incremental predictive effect is assessed separately below. Uniform standardized averaging has retention intervals crossing zero (Table 11).

Estimator choice is separate from the retention effect. Q1 holds the estimator fixed and leaves its choice open. In the broad panel, mean-imputed PC1, the observed-cell rank-one factor, and the affine factor reach 0.652, 0.779, and 0.815. The two successive gains survive Holm correction in this panel. Affine factors have higher prediction point estimates than BoB in both panels. Their comparisons with equal-weight equating are nonsignificant after correction (Table 12). BoB minus equal-weight equating is -0.0016 and +0.0022, with both intervals crossing zero. Thus the retention gains concern the within-estimator pool expansion. A separate comparison on an eligible population tests density weighting's stability effect with affine fitting (Section 4.2). Section C gives objectives, coverage, numerical checks, and 48-round sensitivities.

Description-based hard grouping is a viable alternative. Replacing density weights with reciprocal semantic-cluster sizes also gives positive full-minus-curated effects in both panels. With matched effective mass and fitting conventions, the full-pool difference between BoB and semantic-group weighting is +0.0093 [+0.0011, +0.0183] in the strict panel and +0.0036 [-0.0004, +0.0080] in the broad panel. These marginal intervals show a panel-dependent predictive advantage (Section E.1).

Protocol comparability limits the retention conclusion. A metadata screen removes 38 columns with unresolved evaluation-setting differences. Under the broad panel rules, BoB's retention effect remains positive at +0.0241 over 66 targets, with a marginal 95% CI [0.002, 0.050], but is inconclusive after correction over twenty comparisons (Holm $p = 0.6093$). Only 3 targets survive the strict

Table 4: Ranking sensitivity under complete and partial re-listings. Mean absolute rank displacement is averaged over all 354 WildScores models and then scenarios. Each scenario adds four copies. Coverage is conditional on observing the original column, and all four copies share one mask. The density bandwidth is fixed at its pre-copy value. Scenarios are interventions, not independent sampling trials.

Copy coverage	Scenarios	Equal-weight equating	BoB
Complete	8	1.948	0.625
25%	24	0.684	0.761
50%	24	1.203	0.706
75%	24	1.609	0.657
Existing-column pattern	8	1.089	0.727

panel rules. Because the restriction also changes the suite and fitting population, it does not isolate protocol heterogeneity (Sections D.1 and E.3).

4.2 Re-listing sensitivity and reporting overlap

Complete re-listings test sensitivity to repeated observations. Re-listing adds entries without new measurements. For each of eight WildScores benchmarks, we append four copies of its scores and observation mask, then refit. Mean absolute rank displacement is 5.907 for uniform standardized averaging, 1.948 for equal-weight equating, and 0.626 for density weighting, a 67.9% reduction relative to equal-weight equating. Paraphrased descriptions yield 0.955. Worst-case movements remain 44 and 55 ranks, so stability is approximate. Fixing the pre-copy density bandwidth yields 0.625 for exact copies, close to the re-estimated-bandwidth result (Table 14). The partial-reporting comparisons below use this fixed-bandwidth protocol. Under that protocol, semantic grouping yields mean displacement 0.876 for exact copies and 1.440 for paraphrases, compared with BoB's 0.625 and 0.953. BoB has smaller mean rank displacement in all six declared conditions, with mixed worst-case ordering (Table 21).

Reporting overlap bounds the stability advantage. At 25% copy coverage, density weighting increases mean rank displacement to 0.761 from 0.684 under equal-weight equating (Table 4). It reduces displacement at 50% and 75% coverage and under the tested existing-column reporting pattern. An isolated original and k identical copies each receive unnormalized mass $1/(k+1)$. A model observed on the original and r_m copies therefore receives group mass $(1+r_m)/(k+1)$. With four copies, models observed on all five columns retain mass one, whereas models observed only on the original receive one fifth. Preserving the group's column mass does not preserve every model's observed mass. This holds even with perfect copy identities, and partial columns additionally refit curves on different samples. In the frozen matrix, nearest-description pairs have median directional coverage 36.1%, and 66 of 148 directions have coverage at most 25%. Low overlap is therefore present in the observed corpus (Section E.2).

Initialization and neighboring weights explain residual movement. Exact copies are detectable from the score matrix itself. Deduplicating identical score-and-mask columns while keeping the first listing's description would remove them. We use them as a controlled probe of how weights propagate through the fit. When each original's weight is shared exactly among its copies, rankings still move by 0.150 places on average, and this movement disappears in all sixteen cases once the same weights are also used for initialization. Density weighting additionally changes the weights of neighboring benchmarks, so weighted initialization reduces but does not remove its rank displacement (Section C.2).

Independent measurements can improve precision at fixed weight. A separate simulation with known capability identities (Section C.3) shows that fresh independent measurements, unlike copies, reduce estimation error when each capability's total weight is held fixed.

The stability effect transfers to affine fitting. Applying the strict-panel observation and variance filters once to the unperturbed matrix, without target withholding, retains 127 models and 131 columns. On this population, adding density weights to the affine estimator reduces complete-copy mean absolute rank displacement from 1.041 to 0.220. At 25% nominal copy coverage the direction reverses, from 0.546 without density weighting to 0.567 with it (Table 18). The complete-copy benefit and low-overlap reversal therefore occur in both tested estimators. The comparison excludes sparse models after degenerate affine fits on the full population.

4.3 Controlled mechanism reference on Artificial Analysis

The AA snapshot of 2026-09-03 contains 605 configurations and 14 benchmarks. Because this matrix is denser than WildScores (64% of cells observed), we use it to examine mechanisms with less missingness. Fourteen leave-one-benchmark-out folds predict each target from the remaining columns. We evaluate BoB-Support, the task-conditioned extension described in Appendix F, with two metrics. *Level* is held-out Spearman correlation. *Profile* asks whether a method predicts which models do better or worse on the target than their overall ability suggests, computed after removing a quadratic trend in training-pool ability from predicted and observed ranks. With fully nested parameter selection, BoB-Support reaches level 0.904 and profile 0.483 (Sections H.1.2 and H.6).

Matched controls clarify module roles. Removing density weighting changes profile from 0.483 to 0.473 ($p = 0.121$), while increasing ability shift under re-listing from 0.095 to 0.963 standard deviations (Table 38). Ability shift is the maximum absolute change in θ across models, divided by its pre-copy standard deviation, averaged over 14 four-copy scenarios. Reassigning descriptions in 200 reruns yields mean profile 0.358 and maximum 0.447. These results provide evidence for semantic placement in the full procedure and density weighting for multiplicity robustness. The predictive increment from density weighting is nonsignificant in this comparison.

The published AA Intelligence Index is not recomputed after target withholding and serves as a descriptive comparator. Its profile difference from BoB-Support is nonsignificant (Holm $p = 0.292$). Table 36 and the complete AA controls are reported in Section H.

5 Interpretation and limitations

Evidence retention and listing sensitivity. Expanding the tested curated suite improves held-out ranking under both BoB and equal-weight equating. Affine factors have higher predictive point estimates, and density weighting's incremental predictive intervals cross zero. Density weighting's measured benefit is smaller rank displacement under complete re-listings, including with affine fitting. Relative to semantic grouping, mean displacement is smaller and worst-case comparisons are mixed. These results assign distinct roles to evidence retention, estimator choice, and benchmark weighting.

Protocol comparability and reporting coverage. After metadata restriction, the retention contrast remains positive but is inconclusive after correction. Documented provenance provides traceability, while equating leaves item sets, metrics, and evaluation settings unharmonized. Prediction is evaluated conditionally on reported scores and eligible release groups. Unreported outcomes may follow a different distribution, including through outcome-dependent reporting. Under partial reporting, column-weight conservation leaves each model's observed weight dependent on its mask. Density weighting's stability advantage reverses at 25% copy coverage. Low overlap also occurs among observed semantic neighbors, whose different scores and items place natural ranking effects outside the complete-copy and partial-copy estimands.

Analysis and measurement scope. The fixed-population study followed inspection of earlier outcomes and compares one suite with an expanded pool that changes both quantity and composition. Holm correction covers declared sets of comparisons, while the sequence of result-motivated analysis choices lies outside those corrections. Family-level intervals describe variation across benchmark families in the observed matrix. Individual model abilities are reported as point estimates.

Semantic representation. Global ability is a one-dimensional working summary of aggregate scores. Weights depend on descriptions, one tested encoder series (Qwen3-Embedding), and the

specified density scale. Description neighborhoods represent related content at this resolution, with shared protocols and different domains illustrated by the AA τ pair (Section H.5). The weighting rule specifies a capability emphasis, evaluated here through agreement with reported outcomes and sensitivity to re-listings.

6 CONCLUSION

BoB integrates score equating and semantic weighting while making evidence retention and benchmark weighting explicit design choices. In the fixed-population comparison, expanding the curated suite improves held-out ranking under both equating rules. Complete re-listings show a distinct benefit of density weighting: reduced sensitivity to benchmark multiplicity. Reporting-mask and initialization analyses explain why controlling column weights alone leaves fitted comparisons sensitive to how scores are observed and combined. Reporting overlap and protocol comparability delimit the observed gains. The formulation and interventions distinguish three questions in model comparison: whether extra observations help, whether repeated listings gain weight, and how that weight reaches individual models.

Reproducibility statement

The accompanying artifact contains frozen scores, source records, setting mappings, text inputs, embeddings, splits, and results. Numerical prose and tables are generated from saved outputs. Predictions, selected parameters, and shortlists support reconstruction of rankings and paired comparisons. Tests check target isolation, representative selection, support matching, and agreement with the canonical fit. Records distinguish the original evaluation from exploratory extensions and identify fitting populations, stopping settings, failed fits, and perturbation scenarios. Undefined predictions from failed raw-score affine fits are retained. The appendix specifies the procedures, and the anonymous supplement supplies the score corpus, provenance, and experimental texts.

The use of large language models

Language models are the evaluated systems and also assisted benchmark-record extraction, description and paraphrase construction, demonstration-query canonicalization, prose, and code development. They also assisted refinement of experimental protocols, implementation of the follow-up comparisons, and interpretation of saved results. Source review is partial and documented. Experiments use frozen text and embedding inputs. Deterministic analysis and generation scripts produce numerical results from the retained score matrices. Parity tests and independent reaggregation check the follow-up computations. Sections A.5 and B describe provenance. The anonymous data supplement provides all frozen texts and experimental evidence. The authors are responsible for the paper's content and claims.

References

Artificial Analysis. Artificial analysis: Independent analysis of AI models and API providers. `https://artificialanalysis.ai/`, 2026a. Accessed 2026-09-03.

Artificial Analysis. Artificial analysis intelligence benchmarking methodology. `https://artificialanalysis.ai/methodology/intelligence-benchmarking`, 2026b. Accessed 2026-09-18.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. arXiv:2506.07982, 2025. Preprint.

Damien Berriaud and Roger Wattenhofer. Clone-robust weights in metric spaces: Handling redundancy bias in benchmark aggregation. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026a. doi: 10.65109/MJWJ6521.

Damien Berriaud and Roger Wattenhofer. Breaking the illusion of artificial consensus: Clone-robust weighting for arbitrary metric spaces. arXiv:2602.24024, 2026b. Preprint.

Federico Cassano, John Gouwar, Daniel Nguyen, et al. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7), 2023. doi: 10.1109/TSE.2023.3267446.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. Introduces the Humanity's Last Exam (HLE) benchmark; preprint arXiv:2501.14249.

Jianhao Chen, Chenxu Wang, Gengrui Zhang, Peng Ye, Lei Bai, Wei Hu, Yuzhong Qu, and Shuyue Hu. Learning compact representations of LLM abilities via item response theory. arXiv:2510.00844, 2025. Preprint.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. What are the best systems? new perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Chris Ge, Daria Kryvosheieva, Daniel Fried, Uzay Girit, and Kaivalya Hariharan. Agent psychometrics: Task-level performance prediction in agentic coding benchmarks. arXiv:2604.00594, 2026. Preprint.

Yun He, Di Jin, Chaoqi Wang, et al. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A rosetta stone for AI benchmarks. arXiv:2512.00193, 2025. Preprint.

Blair Hudson. Meta-benchmarks for financial-services LLM evaluation. arXiv:2607.01740, 2026. Preprint.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking: Methods and Practices*. Springer, New York, NY, 3rd edition, 2014.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 648–657, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1062.

Tom Leinster and Christina A. Cobbold. Measuring diversity: The importance of species similarity. *Ecology*, 93(3):477–489, 2012. doi: 10.1890/10-2402.1.

Yang Li. Rethinking predictive modeling for LLM routing: When simple kNN beats complex learned routers. arXiv:2505.12601, 2025. Preprint.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. Introduces the MATH-500 subset of the MATH test set.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Pan Lu, Hritik Bansal, Tony Xia, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326, 2024.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. arXiv:2601.11868, 2026. Preprint; introduces Terminal-Bench 2.0.

Jungsoo Park, Ethan Mendes, Gabriel Stanovsky, and Alan Ritter. Anticipatory evaluation of language models. arXiv:2509.20645, 2025. Preprint; introduces the PRECOG corpus.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2519–2536. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.naacl-long.139.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Do these LLM benchmarks agree? fixing benchmark evaluation with BenchBench. arXiv:2407.13696, 2024b. Preprint.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems 38 (NeurIPS), Datasets and Benchmarks Track*, 2025. Introduces the IFBench benchmark; preprint arXiv:2507.02833.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-Proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.346.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Tommy Sha and Stella Zhao. BenchScope: How many independent signals does your benchmark provide? arXiv:2603.29357, 2026. Preprint.

Freda Shi, Mirac Suzgun, Markus Freitag, et al. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2023.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. arXiv:2504.20879, 2025. Preprint.

Alexander Smola. Submodular benchmark selection. arXiv:2605.02209, 2026. Preprint.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. IRT-Router: Effective and interpretable multi-LLM routing via item response theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15629–15644. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.761.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Martha L. Stocking and Frederic M. Lord. Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2):201–210, 1983. doi: 10.1177/014662168300700208.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. SciCode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, St. Julian's, Malta, 2024. Association for Computational Linguistics.

Fei Wang, Xingyu Fu, James Y. Huang, et al. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yiming Wang, Pei Zhang, Jialong Tang, et al. PolyMath: Evaluating mathematical reasoning in multilingual contexts. In *Advances in Neural Information Processing Systems 38 (NeurIPS), Datasets and Benchmarks Track*, 2025b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024.

Simon Willerton. Spread: A measure of the size of metric spaces, 2015. URL `https://arxiv.org/abs/1209.2300v2`. Version 2.

Jihan Yao, Gantavya Bhatt, Arnav Das, Peter Jin, Ke Bao, Qiaolin Yu, Khushi Bhardwaj, Chang Su, Jialei Wang, Yikai Zhu, Sugam Devare, Damon Mosk-Aoyama, Zhen Dong, Venkat Krishna Srinivasan, Yineng Zhang, Oleksii Kuchaiev, Jiantao Jiao, Banghua Zhu, and Jeff Bilmes. Coresets before score sets: Evaluation-unsupervised prompt subset selection for LLM benchmarks. arXiv:2607.09739, 2026. Preprint.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58984–59002, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176, 2025. Technical report.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. EmbedLLM: Learning compact representations of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (ICLR)*, 2025.

APPENDIX GUIDE

Section A specifies fitting, semantic weights, and the fixed-bandwidth mass calculation. Section B records WildScores construction and the original evaluation. Section C gives fixed-population retention comparisons, re-listing and initialization controls, and the independent-measurement simulation. Section D examines metadata restrictions, partial reporting, and transfer to affine fitting. Section E compares global semantic grouping, measures actual reporting overlap, and audits family-level retention effects under the metadata restriction. Sections F and G give the optional task correction and its sensitivities. Section H provides the AA controls, including fully nested selection. The anonymous supplement contains the score corpus, provenance, frozen texts, and embeddings.

A ADDITIONAL METHOD DETAILS

A.1 FITTING DETAILS

Table 5: Global BoB fitting procedure. The canonical estimator stops after six rounds. Each training pool determines its score matrix and geometry.

Input: eligible normalized scores S, observation mask O, description distances D, and number of rounds $T = 6$.

1. Compute $h_{\text{dens}} = 0.25\,\text{median}_{i<j}\, d_{ij}$, $u_i = 1/\sum_j \exp[-d_{ij}^2/(2h_{\text{dens}}^2)]$, and $v_i = Nu_i/\sum_j u_j$ for the N retained columns.
2. Initialize each θ_m by the mean midpoint percentile of its observed scores, then standardize across models. Set $a_i = 2$, $b_i = 0$, and $w_{mi} = O_{mi}$.
3. For each round $t = 1, \ldots, T$, perform the following updates in order.
   (a) Fit each curve by minimizing $\sum_{m:O_{mi}=1} w_{mi}[s_{mi} - \Phi(a_i(\theta_m - b_i))]^2$ with L-BFGS-B over $\log a_i \in [\log 0.2, \log 20]$ and $b_i \in [-4, 4]$.
   (b) Minimize Equation (2) for each model on the grid $[-3.5, 3.5)$ with step 0.02. At an interior minimum, refine with the three-point parabola, limiting the offset to one grid step. A curvature denominator of magnitude at most 10^{-12} gives zero offset.
   (c) Standardize abilities and transport curves jointly: $\theta \leftarrow (\theta - \mu)/s$, $a \leftarrow sa$, $b \leftarrow (b - \mu)/s$.
   (d) Recompute raw residuals $e_{mi} = s_{mi} - f_i(\theta_m)$ and $\sigma_i = \max(\text{sd}_{m:O_{mi}=1}\, e_{mi}, 10^{-4})$, using the population standard deviation. For $t \geq 2$, update observed-cell weights to $\max(10^{-3}, \min(1, (3/|e_{mi}/\sigma_i|)^2))$. Missing cells retain zero weight.
4. Return θ, a, b, residuals, residual scales, and the final weights. Robust updates after a round enter the next round's fit. The last update also supplies weights for the optional residual correction.

Alternating fit. Initialization averages within-benchmark midpoint percentiles across each model's observed scores, assigning average ranks to ties, then standardizes abilities. The fit alternates for six rounds. Given model abilities, each benchmark's (a_i, b_i) is fitted by weighted least squares on its observed column. Given the curves, each θ_m is fitted over $[-3.5, 3.5)$ in steps of 0.02 using Equation (2), then refined by a parabola through the minimum grid point and its two neighbors. Let μ and s be the mean and standard deviation of the updated abilities. We transform all parameters jointly,

$$\theta'_m = (\theta_m - \mu)/s, \qquad a'_i = sa_i, \qquad b'_i = (b_i - \mu)/s.$$

Then $a'_i(\theta'_m - b'_i) = a_i(\theta_m - b_i)$, so standardization preserves all expected scores. Residual scales and robust weights are computed from this consistent state. Curve-optimization bounds apply before this transport and need not bound the transported parameters.

A.2 INVERSE-DENSITY WEIGHTING AND DUPLICATION DECOMPOSITION

Fix the benchmark descriptions and kernel bandwidth before adding a new point. Write $\rho_i = \sum_j k(d_{ij})$ and $k_i = k(d(x, x_i))$. Since the existing pairwise kernel values are unchanged, adding x

changes each existing density to $\rho_i + k_i$ and gives the new point density $\rho_x = 1 + \sum_i k_i$. Therefore

$$\Delta W = \frac{1}{1+\sum_i k_i} + \sum_i \left(\frac{1}{\rho_i + k_i} - \frac{1}{\rho_i} \right) = \frac{1}{1+\sum_i k_i} - \sum_i \frac{k_i}{\rho_i(\rho_i + k_i)}.$$

Holding the original densities ρ_i fixed, ΔW is decreasing in each k_i. This is a conditional statement about the fixed-bandwidth expression. Re-estimating the bandwidth can also change the original densities and is evaluated separately in Section C.2. If every $k_i = 0$, then $\Delta W = 1$. If x coincides with a member of an isolated co-located group G of size n_G, then $k_i = 1$ and $\rho_i = n_G$ for $i \in G$, giving

$$\Delta W = \frac{1}{1+n_G} - \frac{n_G}{n_G(n_G+1)} = 0.$$

Against one isolated benchmark at distance d, $\rho_1 = 1$ and $k_1 = k(d)$ yield $\Delta W = (1-k)/(1+k)$.

The zero is exact only for an isolated group. If a co-located benchmark has neighbors, its copy also increases their densities and ΔW may be slightly negative. Equivalently, a group of n_G co-located benchmarks holds unnormalized mass

$$\frac{n_G}{n_G + c_G}, \qquad c_G = \sum_{j \notin G} k\big(d(x_G, x_j)\big).$$

This differs from the single-copy value $1/(1+c_G)$ by at most c_G. Treating each original benchmark as a singleton before copying, c_G ranges from 0.004 to 0.180 outside the two close pairs in AA. For the two AIME entries it ranges from 0.482 to 0.532.

Relation to similarity-sensitive diversity. Let $Z_{ij} = k(d_{ij})$ and $p_i = 1/N$. The order-zero diversity of Leinster & Cobbold (2012) is

$$D_0^Z(p) = \sum_i \frac{p_i}{(Zp)_i}.$$

For uniform p, $(Zp)_i = \rho_i/N$, so $D_0^Z(p) = \sum_i 1/\rho_i = W$. This identifies effective mass with an existing similarity-sensitive quantity. We use its individual inverse-density terms for benchmark weighting. The identity does not by itself imply that a refitted ranking is invariant to duplicate listings.

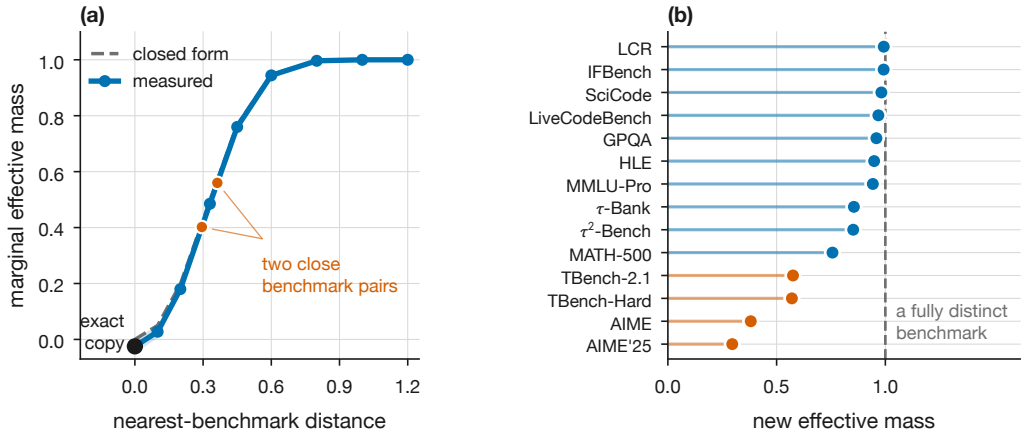

Figure 3: **Effective benchmark mass at a selected semantic scale.** Left: changes in mass across placements, with one standard deviation and the isolated-pair expression. Right: leave-one-out mass in the AA description set. These quantities describe semantic weighting, not independent measurement information.

Why the fitted ranking still moves. We separate three sources under four added copies. Forcing the duplicated group's weight to be exactly invariant and holding item parameters fixed moves no

model. Restoring the actual inverse-density weights while still holding item parameters fixed produces a scenario-averaged maximum displacement of 10.4 ranks from the shared-neighbor $O(c_G)$ term. Refitting characteristic curves and robust cell weights raises the scenario-averaged maximum to 15.9. The corresponding scenario-averaged maximum change in θ is 0.095 standard deviations, while adjacent models are a median 0.0042 apart.

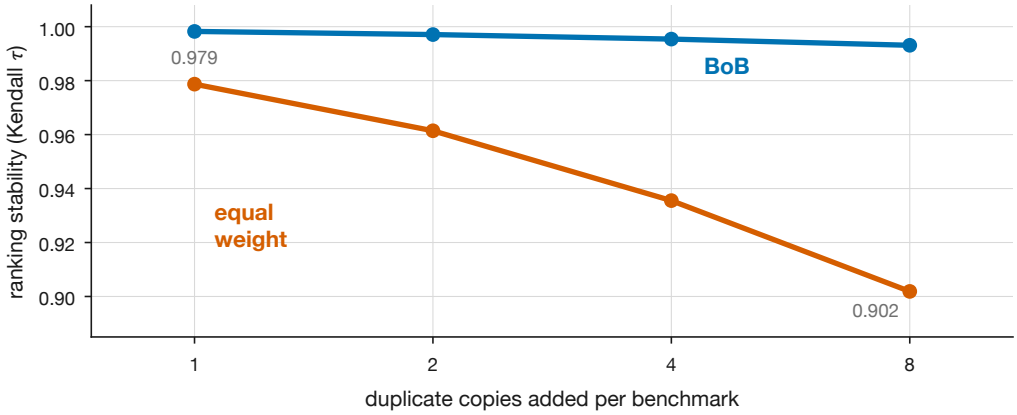

Figure 4: **Inverse-density weighting resists exact benchmark re-listing.** Each point is the mean Kendall τ between the original and perturbed rankings over 14 deterministic scenarios, one per re-listed benchmark. No stochastic interval applies.

A.3 Semantic and score channels

The construction maintains separate semantic and score channels. Figure 1 gives the complete flow in the main method. Figure 2 illustrates score equating across difficulty ranges, Table 6 enumerates the score-side inputs, and Figure 5 expands the factors in the task-conditioned field.

A.3.1 Score weights and information

The robust cell weight is $w_{mi} = \max(10^{-3}, \min(1, (3/|r_{mi}|)^2))$ for observed cells and zero for missing cells. It is updated after the second and subsequent fitting rounds, and enters the next round's fit. Residuals within three standard deviations retain unit weight. Larger residuals are attenuated quadratically. The same weights enter the ability fit, residual field, and support mass.

The mean information is

$$\bar{I}_i = \frac{1}{M} \sum_{m=1}^{M} \frac{\left(a_i \phi(a_i(\theta_m - b_i))\right)^2}{\max(\sigma_i^2, 10^{-6})}, \tag{4}$$

where ϕ is the standard normal density. It records how strongly benchmark i separates the fitted model population relative to its residual scale and is used only for precision weighting inside the residual field.

A.3.2 WHERE SCORES TOUCH THE CONSTRUCTION

Table 6: Uses of model scores in the primary construction. The semantic weights are fixed from descriptions before these score-side quantities enter.

Use of scores	Enters as	Role
Characteristic-curve fit	a_i, b_i, θ_m in Equations (1) and (2)	equates scales and supplies residuals
Per-benchmark residual scale	σ_i in r_{mi} and $\bar{I}_i$	empirical residual standard deviation with numerical floor
Mean item information	$\bar{I}_i$ in Equation (5)	precision-weights the local residual average
Population location and scale	standardization of s_{mi} and θ_m	sets location and scale units
Robust cell weight	w_{mi} in Equations (2) and (5)	limits the influence of one cell's residual
Cross-validation of h	field smoothness	evaluation-side selection with a nested check

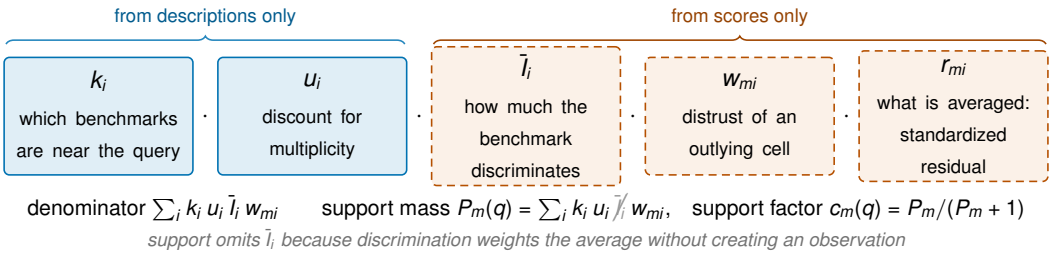

Figure 5: **Semantic and score factors retain distinct roles in the task-conditioned field.** The query kernel and inverse density come from descriptions. Information, robust influence, and residuals come from scores. Information weights direct the residual average. Support mass counts relevant observed evidence through $g_{mi}(q)w_{mi}$.

A.4 TASK-RELEVANCE CONTROLS

We test task relevance by changing only g. These controls share equated scores, residuals, information, robust weights and unit-scale support regularization. *Hard relevance rules* select benchmarks by category, nearest neighbor, top-k, radius or semantic cluster (K-means or agglomerative). They give selected benchmarks unit relevance and select tunable neighborhood sizes on inner training folds. An all-benchmark rule gives every observed benchmark unit relevance. A separate *support-matched* top-k rule chooses k to match the observed support mass of the continuous rule. Because g also enters $P_m^{(g)}(q)$, these comparisons include its effect on correction strength.

A.5 FROZEN TEXT CONSTRUCTION AND EMBEDDING

The fourteen English benchmark descriptions and their paraphrases were generated on 2026-08-29 with `claude-fable-5`. The recorded generation instruction requested a single plain-text English paragraph of 2–4 sentences and roughly 40–80 words identifying the measured domain, ability types, item format, and answer language. A companion paraphrase instruction requested the same meaning in clearly different wording and sentence structure at a similar length, again as one plain-text English paragraph. All fourteen descriptions and their paraphrases were fixed before analysis and are supplied in the anonymous data supplement.

The task demonstrations begin with five user-facing queries. Each was canonicalized once into the description register, with comparable length and facet structure. The supplement includes both the

raw queries and the exact canonical forms used in the experiments. All ten strings were fixed before fitting and evaluation.

All text is embedded with the `Qwen/Qwen3-Embedding-8B` checkpoint through SentenceTransformers. The encoder runs in float16 on an NVIDIA L40S with maximum sequence length 1024 and batch size 8, then stores normalized float32 embeddings. Benchmark descriptions and task queries are encoded symmetrically as documents without an instruction prefix. The reported geometry uses these normalized vectors directly, with no centering or component removal.

The supplement provides the complete WildScores and AA description catalogs in machine-readable and browsable form, together with the frozen vectors and identifiers that align them to score columns. The benchmark source publications describe their underlying evaluations (Rein et al., 2024; Center for AI Safety et al., 2026; Wang et al., 2024; Lightman et al., 2024; Jain et al., 2025; Tian et al., 2024; Pyatkin et al., 2025; Yao et al., 2025; Barres et al., 2025; Merrill et al., 2026).

B WildScores construction and evaluation

Analysis corpus and preparation history. WildScores contains 354 configurations, 316 release groups, 148 benchmark columns, and 5,371 observed cells. Benchmark labels are canonicalized with scale policies and setting provenance. The initial column screen requires at least fifteen settings before the retained-row filter, which requires five observed scores. Each metric applies its additional target-eligibility requirements below.

Score normalization. Canonical scale policies map bounded scores to $[0, 1]$: percentages are divided by 100, fractions retain their scale, and point scores are divided by their documented maximum. The conversion also resolves fractions printed under percentage labels and percentage values tagged as fractions, using the recorded value and its source formatting. Unbounded ratings, unsupported scales, and values outside the normalized range are excluded. For bounded non-accuracy metrics this is a numerical rescaling, not a probability interpretation. Multiple-choice scores retain their reported scale without a chance-floor correction. The supplement records the per-benchmark policies and source-supported corrections.

Source review. An intermediate collection snapshot contained 356 configurations and 5,443 cells. Source review corrected explicit-denominator scale errors for WritingBench and AIME, excluded language-specific appendix results and reasoning-budget ablations that had entered pooled default results, and excluded remaining cells whose source spread exceeded 0.05 on the normalized scale. These corrections define the analysis corpus used throughout the paper. The intermediate snapshot and its analysis outputs are retained as data-preparation history in the accompanying artifact. Source review addresses the documented scale and setting issues. The resulting corpus records scores and evaluation metadata available in the cited developer sources.

Each release group collects a creator/base-model grouping with its listed configurations, using release metadata to define the relationships. Benchmark families combine the named-version groups with documented derivation links, then take connected components for withholding and uncertainty aggregation. MGSM links to GSM8K (Shi et al., 2023), MultiPL-E links HumanEval and MBPP (Cassano et al., 2023), and Multi-IF links to IFEval (He et al., 2024). PolyMath joins the MGSM, HLE, and AIME families through its component sources (Wang et al., 2025b). A visual component connects MMMU, MMBench, AI2D, MathVista, ChartQA, TextVQA, MATH-Vision, HallusionBench, DocVQA, and InfographicVQA through MMStar, DynaMath, MuirBench, and OCRBench (Chen et al., 2024; Lu et al., 2024; Zou et al., 2025; Wang et al., 2025a; Liu et al., 2024). A component's full source families are withheld conservatively. Each version retains its own score column, and family labels never enter the BoB estimator. These documented links address known overlap without certifying complete question-level independence. The accompanying artifact records the mappings, source URLs, settings, and exclusions.

Text and embeddings. Each canonical benchmark has a description and a paraphrase. Existing Qwen3-Embedding-8B outputs provide the semantic vectors. Both frozen text variants and their saved embeddings are included in the anonymous data supplement. Distances are computed from normalized saved vectors. The encoder used document mode with an empty instruction prefix, a

maximum sequence length of 1,024, and 4,096-dimensional outputs. Encoder normalization used float16, with outputs stored in float32 and renormalized in float64 before geometry. Vector and source hashes identify the experimental inputs. The saved vectors provide exact reproducibility. The encoder metadata identifies the model by name, leaving its immutable weight revision unspecified. Follow-up experiments reuse the saved vectors.

Outer and inner splits. For seeds 20260916 and 20260917, a deterministic hash assigns benchmark families to five outer blocks. Every target is withheld once per seed. Three inner family blocks, using the outer seed plus 101, select relevance parameters within the remaining columns. Rows need at least three observed training cells. Rows and columns with fewer than three observations are iteratively removed from each training pool. All such decisions use training missingness. Distance medians, curve fits, residuals, information, and clustering use that pool alone. Tests perturb outer score values and confirm unchanged predictions, inner criteria, selections, and representatives.

Metrics and estimation units. Ranking requires fifteen scored configurations and ten release groups. Configuration scores and predictions are averaged within groups before computing Spearman correlation. Profile removes a quadratic in training-pool ability from ranks. Five-fold target-score calibration withholds whole release groups, requires eight training groups and three distinct predictor values, and fits a monotone isotonic map clipped to the normalized score range. Calibration metrics require ten scored test groups. MAE is expressed in normalized 0–100 score points across the different benchmark metrics.

For each method, metrics are averaged across split seeds within target, then within benchmark family, and then equally across families. Paired comparisons first match method outcomes within the same seed and target. Their percentile intervals use 5,000 family bootstrap resamples. Two-sided family sign-flip tests use 100,000 draws, and Holm adjustment covers eleven continuous-minus-control contrasts separately per metric and fitting-round setting. Ranking correlation and regret at budget three are the primary mechanism and decision metrics. Other budgets and metrics are secondary. These analyses condition on the observed corpus and split seeds. Their intervals describe variation across the declared benchmark families for this score matrix.

B.1 Relevance controls and parameter selection

Relevance controls share characteristic curves, ability, standardized residuals, information, robust cell weights, and $P/(P+1)$ regularization. Continuous relevance is a Gaussian kernel times inverse density. Category, nearest, top-k, radius, and cluster rules use the binary relevance definitions of the AA controls. An all-benchmark residual control gives every observed benchmark unit relevance. θ alone and uniform standardized averaging are global baselines. BoB-Constant retains continuous relevance and replaces support-dependent shrinkage with a constant coefficient.

Continuous bandwidth candidates are $\{0.15, 0.225, 0.30, 0.45, 0.60\}$ times the training median distance. Top-k uses $\{1, 2, 4, 8, 16, 32\}$, radius multipliers are $\{0.15, 0.25, 0.40, 0.55, 0.70, 1.00\}$, and cluster counts are $\{4, 8, 16, 32\}$. K-means uses thirty starts and seed zero, and agglomerative clustering uses average cosine linkage. A query is assigned to the closest normalized training centroid. The constant-coefficient search crosses bandwidth with $c \in \{0.125, 0.25, 0.5, 0.75, 1\}$ and includes θ.

Each tunable rule maximizes family-balanced inner group Spearman on the targets eligible under every candidate. Shortlist outcomes never enter this selection. Support-matched top-k inherits the selected continuous bandwidth, rounds its observed Kish effective-neighbor count, and rescales selected neighbors to reproduce its robust support mass for every model. Assertions verify mass equality to numerical precision. A rule with zero local support returns θ. Complete comparisons use six and forty-eight fitting rounds, with parameter selection repeated separately at each checkpoint.

B.2 Retained-evidence controls

This earlier extension shares evaluation coverage while retaining the original, larger full-pool fitting population. The subsequent fixed-population analysis in Section C refits both pools on identical model rows. The two protocols therefore have distinct predictions even where their eligible ranking-target counts coincide.

Table 7: Earlier retained-evidence extension on common evaluation coverage. The full-pool fitting population remains that of the original evaluation. MAE and regret use 0–100 score points.

Method	ρ	$\Delta\rho$ vs. θ	MAE	R@3	H@3 (%)
Common evaluation coverage: 120 ranking targets, 60 families					
Uniform standardized averaging	0.751	-0.052	5.87	1.99	65.0
AA-v3-derived subset	0.742	-0.060	5.73	1.79	61.4
Mean-imputed PC1	0.631	-0.172	6.60	5.73	25.8
Shrinkage-covariance GLS	0.423	-0.379	8.58	4.94	37.7
BoB global (θ only)	0.802	+0.000	4.93	1.61	71.0
BoB-Support	0.811	+0.008	4.84	1.68	71.7
BoB-Constant	0.816	+0.013	4.74	1.53	74.0

The retained-evidence extension uses the reviewed matrix and saved outer-fold predictions at six rounds. The curated rule is a fixed ten-column adaptation of the historical AA Index v3.0 selection: MMLU-Pro, GPQA Diamond, HLE text-only, LiveCodeBench v6, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, and τ^2-Bench Telecom. AA's official version history records the suite's evolution (Artificial Analysis, 2026b). HLE text-only and LiveCodeBench v6 specify the release-column mapping. Other versions and tool-use settings remain separate columns outside this subset. This control uses developer-reported scores and equal standardized weights. The published AA Index has its own evaluation and weighting rules.

Each outer block removes the complete source-lineage families from every method, including the curated subset. Training-only row and column filters are identical to those used by the original full-pool fit. After observed training-column standardization, the curated rule averages whichever of its remaining columns a model reports. A model with zero such observations has an undefined prediction. The ten folds retain six to ten variable suite columns. The common-coverage comparison intersects finite predictions before inspecting target values and retains the original training-selected shortlist representatives. Ranking requires a defined correlation from every method on the same configurations, release groups and targets. Calibration and shortlist comparisons apply their own shared eligibility rules.

PC1 fills missing standardized scores with zero, centers columns, and takes the first right singular vector. Its sign is oriented by training-row means. The main covariance control fits a Ledoit–Wolf covariance to the same mean-imputed standardized score matrix, adds $0.05I$, and solves for weights proportional to $C^{-1}\mathbf{1}$. Predictions normalize these weights over each model's observed cells. An absolute denominator at most 10^{-9} is undefined. Shrinkage is estimated from training scores alone. These are specified score-structure baselines with their own missingness conventions. The comparison with BoB evaluates the complete estimation procedures.

We also retain the AA pairwise-correlation GLS rule. It row-centers observed standardized scores, computes pairwise correlations, sets unidentified cross-correlations to zero, and adds ridge 0.05. On the sparse release matrix, this gives non-positive-definite matrices despite the ridge. The positive-definite control was added after observing this limitation. A first sensitivity that shrank mean-imputed, row-centered residuals gives uniform weights algebraically, because each residual row sums to zero. The direct-score covariance control above supplies a distinct weighting comparison. All three outcomes and the result-motivated amendments are retained.

The extension reports all targets, exact curated-suite members, and other targets separately. The ranking counts are 120 targets in 60 families overall, 9 in 9 families inside the suite, and 111 in 57 families outside. A lineage family can contribute to both descriptive strata. Paired differences average seeds within targets and targets within families. Intervals use 5,000 family bootstrap draws. Two-sided sign-flip tests use 100,000 draws. For each metric, Holm correction covers twelve comparisons in each of three strata: seven rules versus θ, Support versus four new controls, and full-pool uniform versus curated mean. Ranking is primary, while score calibration and shortlists are secondary. This extension follows the initial results and has its own declared comparisons and coverage.

Table 8: Support settings on WildScores. Canonical denotes the reference procedure used in the main comparisons.

Support setting	Seeds	h grid	λ	α	ρ
Canonical Support	2	H_5	1	1	0.8232
Fixed Support	2	$\{0.45\}$	1	1	0.8210
Initial bandwidth search	2	H_3	1	1	0.8246
Fixed, first seed	1	$\{0.45\}$	1	1	0.8248
Initial joint search	2	H_3	L_4	A_3	0.8242
Expanded joint search	2	H_5	L_5	A_5	0.8244

All rows use six rounds, density multiplier 0.25, and the reviewed corpus. $H_3 = \{0.30, 0.45, 0.60\}$ and $H_5 = \{0.15, 0.225, 0.30, 0.45, 0.60\}$ multiply the training distance median. $L_4 = \{0.25, 0.5, 1, 2\}$, $L_5 = \{0.125, 0.25, 0.5, 1, 2\}$, $A_3 = \{0.25, 1, 4\}$ and $A_5 = \{0.0625, 0.125, 0.25, 1, 4\}$. Joint searches include θ. The canonical row selects h by inner family-balanced ranking. The earlier extension in Table 7 applies the same saved canonical predictions to common evaluation coverage, giving 0.811.

Table 9: Common-coverage comparisons inside and outside the fixed curated suite. MAE uses normalized 0–100 score points.

	All		Inside suite		Outside suite	
Method	ρ	MAE	ρ	MAE	ρ	MAE
Uniform standardized averaging	0.751	5.87	0.683	10.12	0.752	5.71
AA-v3-derived subset	0.742	5.73	0.786	8.62	0.738	5.64
Mean-imputed PC1	0.631	6.60	0.583	9.97	0.631	6.50
Pairwise-correlation GLS	0.389	8.72	0.351	12.66	0.393	8.57
Shrinkage-covariance GLS	0.423	8.58	0.374	12.49	0.418	8.50
BoB global (θ only)	0.802	4.93	0.837	7.12	0.801	4.84
BoB-Support	0.811	4.84	0.847	6.98	0.809	4.75
BoB-Constant	0.816	4.74	0.863	6.69	0.813	4.68

Table 10: Selected paired ranking contrasts in the retained-evidence extension. Holm correction covers all 36 declared ranking comparisons.

Targets	Paired comparison	$\Delta\rho$ [95% CI]	Holm p
All	AA-v3-derived subset − BoB global (θ only)	-0.060 [-0.081, -0.038]	< 0.001
All	Mean-imputed PC1 − BoB global (θ only)	-0.172 [-0.243, -0.107]	< 0.001
All	Shrinkage-covariance GLS − BoB global (θ only)	-0.379 [-0.435, -0.326]	< 0.001
All	BoB-Support − BoB global (θ only)	+0.008 [+0.004, +0.014]	0.004
All	BoB-Support − AA-v3-derived subset	+0.068 [+0.046, +0.090]	< 0.001
All	Uniform standardized averaging − AA-v3-derived subset	+0.008 [-0.021, +0.039]	0.779
Inside	AA-v3-derived subset − BoB global (θ only)	-0.051 [-0.075, -0.025]	0.058
Inside	Mean-imputed PC1 − BoB global (θ only)	-0.254 [-0.382, -0.155]	0.057
Inside	Shrinkage-covariance GLS − BoB global (θ only)	-0.463 [-0.552, -0.374]	0.057
Inside	BoB-Support − BoB global (θ only)	+0.010 [+0.003, +0.021]	0.058
Inside	BoB-Support − AA-v3-derived subset	+0.061 [+0.031, +0.090]	0.058
Inside	Uniform standardized averaging − AA-v3-derived subset	-0.104 [-0.157, -0.053]	0.057
Outside	AA-v3-derived subset − BoB global (θ only)	-0.063 [-0.084, -0.040]	< 0.001
Outside	Mean-imputed PC1 − BoB global (θ only)	-0.170 [-0.250, -0.100]	< 0.001
Outside	Shrinkage-covariance GLS − BoB global (θ only)	-0.382 [-0.442, -0.323]	< 0.001
Outside	BoB-Support − BoB global (θ only)	+0.008 [+0.004, +0.014]	0.008
Outside	BoB-Support − AA-v3-derived subset	+0.071 [+0.048, +0.093]	< 0.001
Outside	Uniform standardized averaging − AA-v3-derived subset	+0.014 [-0.018, +0.045]	0.779

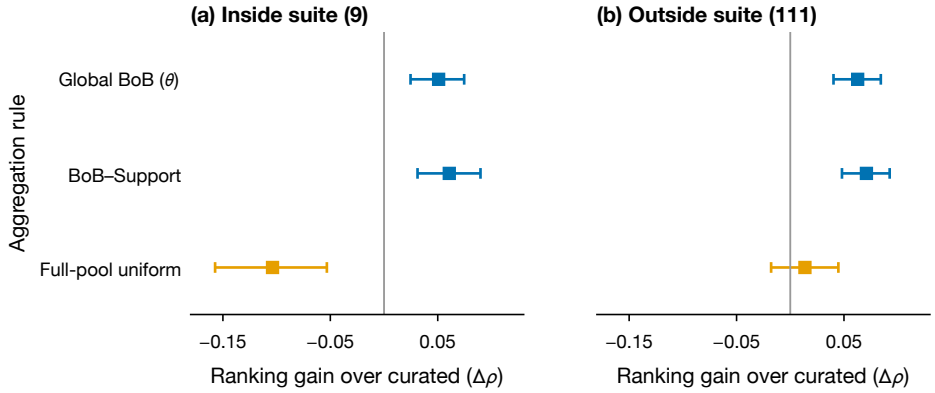

Figure 6: Earlier extension: paired ranking effects relative to the fixed curated mean, with shared evaluation coverage. Intervals use 5,000 family bootstrap resamples. Corrected tests are in Table 10.

B.3 Retrospective shortlist design

Before inspecting target scores, each outer fit selects one configuration per release group by the number of observed training scores, breaking ties by configuration ID. For each query, the common candidate pool retains those representatives with an observed target score and finite predictions under every compared rule. At least ten candidates are required. Representatives are fixed before target-availability filtering. A missing target score removes that representative from the candidate pool.

For $k \in \{1, 3, 5\}$, predictions order the candidates, with configuration ID resolving prediction ties. Regret is

$$R_k(q) = 100 \left[\max_{m \in C_q} s_{mq} - \max_{m \in A_k(q)} s_{mq} \right],$$

where C_q is the common observed pool and $A_k(q)$ is the nominated set. Hit records whether the set contains any candidate tied for the maximum observed score. For larger budgets, these outcomes assume evaluating every nominated model. Their reference population consists of candidates with observed target outcomes. Candidate IDs, scores, and predictions are retained for direct reproduction of every decision metric.

C Evidence retention and benchmark multiplicity

This exploratory analysis was designed after inspecting the retained-evidence extension in Section B.2. The factorial protocol, the listing and simulation protocol, and their diagnostic amendments specify successive result-motivated stages. All settings and outcomes are retained. The six-round schedule and original evaluation population are retained in Table 25. All fits incorporate the prediction-preserving scale transport described in Section A.1.

C.1 Populations, estimators and inference

The follow-up uses the same 354-by-148 reviewed matrix, fixed ten-column AA-v3-derived suite, two split seeds and five outer source-family blocks. Every fit excludes the entire outer target families. Training-only filtering recursively requires three observed models per column and removes constant columns. The stricter panel also requires three curated observations per model. The lower-support sensitivity allows one curated observation after the original full-pool three-observation screen. Within a panel, curated and full fits use exactly the same model rows. The full pool then includes all training columns supported by these rows. The strict panel retains 30–127 models per

fold and evaluates 50 ranking targets in 29 families. The sensitivity retains 207–300 models and evaluates 120 targets in 60 families. All forty model–benchmark observation graphs are connected.

The original six-round fit uses density multiplier 0.25 and each pool's own training-distance median. Retaining columns thus changes both available score evidence and the description geometry used by the same estimator. A separate 48-round comparison repeats the procedure, with both stopping settings reported. The observed standardized mean averages available cells. PC1 replaces missing standardized entries by zero and uses the first singular component, oriented by observed training-row means.

For the observed-cell factors, let z_{mi} be the score standardized using observed training-column means and standard deviations. The affine model minimizes

$$\sum_{(m,i)\in\Omega} \left[z_{mi} - (c_i + \ell_i t_m)\right]^2,$$

where Ω contains observed training cells. The masked rank-one model fixes $c_i = 0$. Alternating exact least-squares updates estimate latent scores, loadings and, for the affine model, intercepts. Normalizing the factor at each iteration preserves its reconstruction. Three deterministic starts use observed row means, mean-imputed PC1 and a fixed random seed. Training reconstruction loss selects the start. Relative loss tolerance is 10^{-9} with an initial 2,000-iteration cap. Factors have unrestricted loadings, whereas the canonical characteristic curves require positive slopes. The full-pool affine fits have zero to three negative loadings per fold. Loss scales, robust weighting and parameter constraints also differ. These are complete-estimator comparisons.

Source-family weighting assigns inverse training-family size to each column, using lineage metadata as an explicitly assisted control. Family weights are recomputed within each retained pool. They share total family influence while allowing different members of that family.

All methods share eligible configurations and targets within a panel and stopping setting. Ranking requires at least fifteen configurations and ten release groups. Seeds are averaged within targets, then targets within families. Each family has equal weight. Paired intervals use 5,000 family bootstrap draws, and two-sided sign-flip tests use 100,000 draws. The factorial set has 84 tests: seven contrasts across two support panels, two stopping settings and three target strata. Six estimator contrasts form a separate 72-test set. Holm correction is applied within each set of comparisons. Intervals are marginal 95% intervals. Shortlist metrics are secondary and descriptive, using identical training-selected representatives.

Table 11: Six-round factorial contrasts on all eligible targets. Holm p covers 84 contrasts across support, stopping and target strata.

Suite support	Comparison	$\Delta\rho$	95% CI	Holm p
≥ 3	Mean (full) − Mean (curated)	+0.0076	[-0.0185, +0.0365]	1.0000
≥ 3	Equal equating (full) − Equal equating (curated)	+0.0367	[+0.0208, +0.0529]	0.0050
≥ 3	Density equating (full) − Density equating (curated)	+0.0350	[+0.0183, +0.0520]	0.0130
≥ 3	Equal equating (curated) − Mean (curated)	-0.0159	[-0.0317, -0.0003]	1.0000
≥ 3	Density equating (curated) − Equal equating (curated)	+0.0000	[+0.0000, +0.0001]	1.0000
≥ 3	Equal equating (full) − Mean (full)	+0.0132	[-0.0074, +0.0337]	1.0000
≥ 3	Density equating (full) − Equal equating (full)	-0.0016	[-0.0102, +0.0076]	1.0000
≥ 1	Mean (full) − Mean (curated)	+0.0226	[-0.0035, +0.0492]	1.0000
≥ 1	Equal equating (full) − Equal equating (curated)	+0.0342	[+0.0180, +0.0496]	0.0075
≥ 1	Density equating (full) − Density equating (curated)	+0.0364	[+0.0186, +0.0540]	0.0107
≥ 1	Equal equating (curated) − Mean (curated)	+0.0227	[+0.0091, +0.0348]	0.0598
≥ 1	Density equating (curated) − Equal equating (curated)	-0.0000	[-0.0000, +0.0000]	1.0000
≥ 1	Equal equating (full) − Mean (full)	+0.0343	[+0.0122, +0.0566]	0.2349
≥ 1	Density equating (full) − Equal equating (full)	+0.0022	[-0.0023, +0.0067]	1.0000

Table 12: Six-round estimator contrasts on all eligible targets. Holm p covers the separate 72-test set. Full and curated denote column pools.

Suite support	Comparison	$\Delta\rho$	95% CI	Holm p
≥ 3	Equal equating (full) − PC1 (full)	+0.0877	[+0.0494, +0.1294]	0.0043
≥ 3	Equal equating (full) − Masked factor (full)	+0.0098	[-0.0062, +0.0271]	1.0000
≥ 3	Equal equating (full) − Affine factor (full)	-0.0084	[-0.0257, +0.0063]	1.0000
≥ 3	Masked factor (full) − PC1 (full)	+0.0779	[+0.0429, +0.1164]	0.0071
≥ 3	Affine factor (full) − Masked factor (full)	+0.0182	[+0.0044, +0.0324]	0.6403
≥ 3	Density equating (full) − Family equating (full)	-0.0131	[-0.0265, -0.0017]	1.0000
≥ 1	Equal equating (full) − PC1 (full)	+0.1474	[+0.0875, +0.2138]	0.0007
≥ 1	Equal equating (full) − Masked factor (full)	+0.0205	[+0.0040, +0.0366]	0.6919
≥ 1	Equal equating (full) − Affine factor (full)	-0.0156	[-0.0300, -0.0037]	0.5072
≥ 1	Masked factor (full) − PC1 (full)	+0.1269	[+0.0700, +0.1906]	0.0014
≥ 1	Affine factor (full) − Masked factor (full)	+0.0361	[+0.0200, +0.0549]	0.0020
≥ 1	Density equating (full) − Family equating (full)	-0.0045	[-0.0104, +0.0015]	1.0000

Suite membership and numerical scope. Nine eligible ranking targets belong to the suite in each panel. Outside-suite density-equating retention gains are +0.0373 over 41 targets in 24 families and +0.0379 over 111 targets in 57 families, with Holm $p = 0.0673$ and 0.0136, respectively. These strata describe one fixed suite under the stated reporting population. The forty-eight-round comparison is in Table 13.

Table 13: Fixed-population sensitivity at 48 equating rounds. Factor and mean estimators are unchanged. † uses source-lineage metadata.

	≥ 3 suite observations			≥ 1 (sensitivity)		
Method	Curated	Full	Δ	Curated	Full	Δ
Existing aggregation rules						
Uniform standardized averaging	0.819	0.827	+0.008	0.742	0.765	+0.023
Mean-imputed PC1	0.772	0.752	-0.019	0.624	0.652	+0.028
Observed-cell rank-one factor	0.819	0.830	+0.012	0.721	0.779	+0.058
Observed-cell affine factor	0.823	0.849	+0.026	0.764	0.815	+0.051
BoB and ablations						
BoB global (θ, density)	0.801	0.834	**+0.032**	0.766	0.799	**+0.033**
− density (equal-weight equating)	0.802	0.834	+0.032	0.766	0.798	+0.032
Metadata-assisted control						
density → source-family weights†	0.797	0.850	+0.054	0.764	0.807	+0.043

Convergence checks extend one capped standardized-affine fit to convergence without changing its held-out ranking metrics. A separate raw-score affine diagnostic changes the loss scaling and encounters two degenerate fitting pools, whose predictions remain undefined. Convergence summaries, raw-score comparisons and reporting-support strata are retained in the artifact.

C.2 Listing robustness and initialization

The full reviewed matrix supplies eight benchmarks selected by observed coverage across eight categories. Each receives four score copies, with exact or stored paraphrased descriptions. The sixteen scenarios compare uniform standardized averaging, equal-weight equating, semantic density with refitted or fixed bandwidth, source-family weights, and known-copy-ID weights. The weighted initialization diagnostic applies the same column weights to the initial percentile-rank mean. Its default unweighted path exactly matches the canonical implementation. Weighted initialization is a diagnostic variant. Table 14 reports every tested variant.

Table 14: Rank displacement after four repeated listings. Mean averages absolute model displacement within each of eight scenarios, then scenarios. Worst is the largest single-model displacement across the eight scenarios.

	Exact		Paraphrase	
Method	Mean	Worst	Mean	Worst
Uniform	5.907	127	5.907	127
Equating, equal weights	1.948	73	1.948	73
Density, refitted bandwidth	0.626	44	0.955	55
Density, fixed bandwidth	0.625	44	0.953	55
Density, weighted initialization	0.569	42	0.917	55
Density, fixed bandwidth + weighted initialization	0.567	42	0.912	55
Source-family weights	1.011	51	1.011	51
Source-family weights + weighted initialization	0.978	51	0.978	51
Known-copy ID weights	0.150	5	0.150	5
Known-copy ID weights + weighted initialization	0.000	0	0.000	0

Known-copy weights share one total weight among each original and its copies. With ordinary initialization, exact-copy mean movement is 0.150 places. Applying the same weights during initialization gives zero rank movement in all sixteen cases and maximum latent-score change about 3.4×10^{-8}. This control has exact copy identities. Source-family labels are coarser and may group different benchmarks, so within-family composition can still change.

Soft density also couples neighboring weights. At fixed bandwidth, let ρ_q be the original density at benchmark q. Adding k co-located copies makes their combined unnormalized mass

$$M_q(k) = \frac{k+1}{\rho_q + k}, \qquad M_q(k) - M_q(0) = \frac{k(\rho_q - 1)}{\rho_q(\rho_q + k)}.$$

An isolated original has $\rho_q = 1$ and constant group mass. Neighboring entries give $\rho_q > 1$, so this group's mass increases while neighbors' inverse densities change, followed by global normalization. This identity explains a boundary of approximate multiplicity robustness. Fixed bandwidth and weighted initialization give modest reductions in the semantic-density perturbations, consistent with remaining neighborhood and refitting effects.

C.3 Independent measurements under fixed capability weights

The frozen source records contain aggregate scores and provenance, with item responses outside their resolution. The following simulation studies a weighting mechanism with explicitly independent measurement noise. It evaluates weighted score averages, separately from end-to-end BoB fitting.

Thirty independent worlds each contain 100 models and nine latent capability probabilities $p_{md} = \Phi(0.8g_m + 0.6a_{md})$, with independent standard-normal g_m and a_{md}. Each world has thirty measurement repetitions with 80-item Binomial observations. Starting with eight capabilities, additions $k \in \{0, 1, 2, 4, 8, 16\}$ either copy the original score on capability one, provide independent samples from that capability, or measure the ninth capability. The first two conditions target the mean latent probability of the original eight capabilities. The third targets all nine, including at $k = 0$. Exact inverse-count weighting uses the true capability identities. Soft density uses nine equally spaced capability positions on a circle, shortest angular distances and bandwidth 0.4 radians. Five thousand bootstrap draws across worlds form intervals after averaging measurement repetitions within worlds.

Table 15: Simulation MSE after four additions, in units of 10^{-4}. The reference target is the equal-capability latent probability mean.

Four additions	Uniform	Exact count	Soft density
Score copies	34.962	2.630	4.155
Fresh samples, same capability	31.928	2.368	3.675
Fresh samples, new capability	18.904	2.142	3.248

With fixed capability weights w_d, averaging independent measurements within each capability gives

$$\widehat{T}_m = \sum_d w_d \bar{s}_{md}, \qquad \mathrm{Var}(\widehat{T}_m) = \sum_d w_d^2 \sigma_{md}^2 / n_d,$$

assuming measurement errors are independent within and across capabilities. Exact-count weighting reduces MSE by 10.0% after four fresh samples while keeping that capability's total weight at $1/8$. Copies preserve its original error. Soft density also benefits from independent samples relative to the same number of copies, but changing neighborhood mass introduces composition bias under this equal-capability target. Its error can therefore increase relative to the starting pool. The known capability identities and chosen geometry define the scope of this mechanism study.

D PROTOCOL COMPARABILITY AND PARTIAL REPORTING

These exploratory analyses test two conditions behind the main results: whether retained scores describe comparable measurements, and whether repeated columns preserve their observation masks. Protocols, exclusion decisions and complete outcomes are available in the accompanying artifact.

D.1 METADATA-RESTRICTED EVIDENCE

The frozen metadata identifies unresolved versions or item windows, scoring levels, voting aggregation, language sets, tool availability and judge or environment differences. Generic LiveCodeBench mixes date windows, IFEval does not consistently distinguish its four scoring levels, and MMMLU varies its language set. SciCode includes both main-problem and sub-problem scores. A conservative whole-column screen, fixed before evaluating predictions, removes 38 columns and 1,750 cells, leaving 110 columns and 3,621 cells. This is a metadata-supported restriction, not certification of identical prompts, shot counts, generation budgets or agent harnesses.

The curated suite retains five components: AA-LCR, HLE text-only, IFBench, MMLU-Pro and Terminal-Bench Hard. Each outer fold applies the restriction before fixing common model rows for full and curated fits. Three of twenty panel–seed–fold combinations become ineligible under the original support thresholds. The three-observation rule leaves only 3 targets in 3 families, which is insufficient to substantiate a retention advantage. The one-observation rule retains 52–199 fitting models per completed fold and 66 targets in 43 families. Some targets consequently contribute only one seed.

Table 16: Retention after the metadata restriction. Family-mean Spearman on 66 targets in 43 families under the one-observation rule. Full and curated fits share model rows. Marginal intervals use 5,000 family bootstrap draws, and Holm correction covers all twenty declared retention comparisons.

Estimation	Curated	Full	$\Delta\rho$	95% CI	Holm p
Standardized mean	0.754	0.760	+0.006	[-0.032, +0.040]	1.000
Equal equating	0.791	0.813	+0.022	[-0.002, +0.050]	1.000
Density equating	0.791	0.815	+0.024	[+0.002, +0.050]	0.609
Affine factor	0.795	0.802	+0.007	[-0.018, +0.031]	1.000
Density affine	0.795	0.804	+0.009	[-0.013, +0.031]	1.000

The restricted density-equating gain is +0.0241 (Holm $p = 0.6093$). Thus the broader corpus's significant retention gain does not remain conclusive under this restriction. On these same fitting rows, excluding only the target columns changes original-full-pool density correlation from 0.796 to 0.794. Excluding those columns from training as well gives 0.815. Training and target exclusion therefore have distinct effects. Because the supported suite and population also change, the comparison does not isolate the causal effect of heterogeneous measurements.

D.2 PARTIAL COPIES AND ESTIMATOR TRANSFER

For each of the same eight benchmarks, we append four copies retaining 25%, 50%, or 75% of the models observed on the original column, using three fixed random seeds. These percentages are

conditional copy coverages, not fractions of all model-benchmark cells. An additional mask follows an existing column's reporting pattern, selected for overlap closest to 50% using missingness alone. All original observations remain, and the four copies within a scenario share one mask. Including complete-copy references gives 88 scenarios. The primary comparison fixes the original density bandwidth. This design isolates selected reporting-overlap conditions but does not estimate their frequency among naturally related benchmarks.

Table 17: Mean absolute rank displacement on all 354 models. Partial-copy fractions refer to models with an observed original score. Cases count perturbation scenarios, which are stress tests rather than independent trials.

Copy coverage	Cases	Uniform	Equal equating	Density
Full	8	5.907	1.948	0.625
25%	24	1.855	0.684	0.761
50%	24	3.418	1.203	0.706
75%	24	4.681	1.609	0.657
Reporting pattern	8	3.090	1.089	0.727

Density reduces mean movement at 50% and 75% coverage, but at 25% its 0.761 displacement exceeds equal-weight equating's 0.684. Among unexposed models, density changes the within-group ordering by 0.668 places and original-cell fitted predictions by 0.096 score points on average. The effect is therefore not solely other models passing them in rank.

The limitation arises even with perfect copy identity. An isolated original and k copies each receive mass $1/(k+1)$. A model observed on the original and r_m copies has group mass $(1+r_m)/(k+1)$. With four copies, exposed models retain mass one while models observed only on the original receive one fifth. Column-level mass conservation does not conserve each model's observed evidence. Partial columns also refit curves on different samples.

Transfer to affine factors. A density-weighted affine control minimizes $\sum_{(m,i)\in\Omega} u_i[z_{mi} - (c_i + \ell_i t_m)]^2$. It shares the equal-weight affine estimator's observed standardized scores, three starts and loss tolerance, with a 20,000-iteration cap for both. Equal weights reproduce the existing solver in a parity test. Both affine fits encounter degenerate loading updates on the full 354-model matrix. A subsequent diagnostic therefore fixes the population using the existing three-curated-observation rule, yielding 127 models and 131 columns, then intersects the original masks with those rows.

Table 18: Affine-factor rank movement on the supported 127-model population. Both columns use the same finite scenarios and fixed original bandwidth. Three masks have no copied-score variation and one further equal-affine fit fails, leaving 84 matched scenarios.

Copy coverage	Cases	Affine	Density affine
Full	8	1.041	0.220
25%	23	0.546	0.567
50%	22	0.709	0.437
75%	24	0.905	0.289
Reporting pattern	7	1.030	0.531

For complete copies, density reduces affine movement from 1.041 to 0.220 places, so the stability benefit extends beyond nonlinear equating. At 25% nominal coverage, movement is instead 0.567 with density versus 0.546 without it. Observation-mask overlap is therefore a condition for the observed advantage in both estimators. This diagnostic does not establish affine stability for the excluded sparse models. Worst-case movements, unexposed-model diagnostics, bandwidth sensitivities and all failed fits remain available in the artifact.

E GLOBAL GROUPING AND OBSERVED SUPPORT DISTRIBUTIONS

These bounded exploratory analyses address three distinct questions: the choice of a global semantic weighting rule, overlap among naturally neighboring descriptions, and the distribution of the metadata-restricted retention effect. The grouping protocol was fixed before fitting this comparator, after the earlier results were known. The remaining analyses audit existing data and saved outcomes. Their evidence is descriptive and exploratory, rather than independent confirmation.

E.1 EFFECTIVE-MASS-MATCHED SEMANTIC GROUPING

We replace the global item weights v_i using the same frozen descriptions and angular distances. Complete-linkage clustering groups columns by their maximum pairwise distance. Within each training pool, we compute the original bandwidth $h = 0.25\ \mathrm{median}_{i<j}\, d_{ij}$ and soft effective mass $W = \sum_i \rho_i^{-1}$. Cutting the hierarchy at $G = \min(N, \max(1, \lfloor W + 1/2 \rfloor))$ clusters gives each column unnormalized weight $u_i = 1/|c(i)|$, where $|c(i)|$ is its cluster size. We normalize these weights as in Section 3.2. Each cluster has unit total mass. Frozen column order resolves ties deterministically under the recorded clustering implementation.

On all 148 columns, $W = 104.43$ gives $G = 104$. A complete-linkage cut at the Gaussian kernel's half-height instead gives 141 clusters, illustrating why an uncalibrated distance threshold changes granularity. Matching G to W aligns rounded total mass, while neighborhood sizes and weight distributions can still differ. The comparator is calibrated from soft density, so the experiment evaluates these two specified rules and does not isolate continuity alone.

Both methods independently refit characteristic curves and robust cell weights with the canonical six rounds, observed-cell objective, support filters and unweighted percentile-rank initialization. A paired sensitivity uses each method's item weights during initialization for both methods. Neither comparison changes the optional residual operator g. Training populations, seeds, family withholding and target eligibility reproduce the original fixed-population panels, with no additional attrition. Each training pool determines its own h, W and G.

Table 19: Soft density minus global semantic grouping in held-out Spearman. The stricter panel has 50 targets in 29 families and the broader panel has 120 targets in 60 families. Intervals are marginal 95% family bootstrap intervals after averaging seeds within targets and targets within families.

Support	Pool	Initialization	$\Delta\rho$	Marginal 95% CI
≥ 3	Curated	Unweighted	+0.000027	[+0.000000, +0.000082]
≥ 3	Full	Unweighted	+0.009314	[+0.001116, +0.018338]
≥ 3	Curated	Weighted	+0.000027	[+0.000000, +0.000082]
≥ 3	Full	Weighted	+0.010414	[+0.001658, +0.020087]
≥ 1	Curated	Unweighted	-0.000003	[-0.000019, +0.000007]
≥ 1	Full	Unweighted	+0.003592	[-0.000387, +0.007970]
≥ 1	Curated	Weighted	-0.000003	[-0.000019, +0.000007]
≥ 1	Full	Weighted	+0.003435	[-0.000409, +0.007739]

The full-pool contrast favors density more under the stricter support rule, with a positive marginal interval. The broader panel's interval crosses zero. Curated-pool differences are small, and weighted initialization preserves these patterns. The intervals are exploratory, not simultaneous intervals or equivalence tests. Grouping also benefits from retaining the full pool (Table 20), so evidence-retention gains are compatible with both semantic weighting rules.

Table 20: Full-minus-curated retention with semantic-group weights. Marginal intervals use the same family bootstrap as Table 19.

Support	Initialization	$\Delta\rho$	Marginal 95% CI	Targets/families
≥ 3	Unweighted	+0.0257	[+0.0086, +0.0443]	50/29
≥ 3	Weighted	+0.0246	[+0.0071, +0.0434]	50/29
≥ 1	Unweighted	+0.0328	[+0.0129, +0.0524]	120/60
≥ 1	Weighted	+0.0330	[+0.0131, +0.0526]	120/60

Listing stability. We reuse all eight original listing targets, the four added columns, the complete-copy masks and every partial-copy mask. Eight full paraphrase scenarios bring the total to 96. The original bandwidth remains fixed, while the expanded descriptions determine new W, G and cluster assignments. Copy identities never supply group membership. Every variant receives its own baseline fit. Actual exposed row identifiers and all per-scenario outcomes accompany the results.

Table 21: Listing stability for the two global weighting rules. Each entry is mean absolute rank displacement, with the worst individual displacement in parentheses. Means average models, then scenarios within each condition. Weighted initialization uses the corresponding rule's own weights for both methods. Scenarios are interventions, not iid trials.

		Unweighted initialization		Weighted initialization	
Condition	Cases	Density	Grouping	Density	Grouping
Complete exact	8	0.625 (44)	0.876 (45)	0.567 (42)	0.809 (41)
Complete paraphrase	8	0.953 (55)	1.440 (94)	0.912 (55)	1.406 (90)
25% coverage	24	0.761 (42)	0.799 (47)	0.758 (41)	0.787 (45)
50% coverage	24	0.706 (51)	0.810 (51)	0.697 (49)	0.778 (41)
75% coverage	24	0.657 (52)	0.862 (51)	0.615 (50)	0.810 (46)
Reporting pattern	8	0.727 (50)	0.869 (52)	0.706 (50)	0.866 (49)

Density has smaller mean displacement in all six conditions under both initializers. Worst-case comparisons are mixed, including a smaller worst exact-copy movement for grouping under weighted initialization. Both rules have larger mean movement than equal-weight equating at 25% coverage (Table 4). Even within a fixed hard cluster, adding k copies of one of n original measurements changes its share from $1/n$ to $(k+1)/(n+k)$. Conserving total cluster mass therefore permits changes in within-cluster measurement influence, in addition to changes from partitioning and fitting.

E.2 Reporting overlap among semantic neighbors

Let O_i be the configurations reporting column i in the frozen 354-model matrix. We calculate directional coverage $C(j \mid i) = |O_i \cap O_j|/|O_i|$ and Jaccard overlap $J(i,j) = |O_i \cap O_j|/|O_i \cup O_j|$ on this configuration-level mask. Observations are not unioned across release groups. Each column contributes one nearest-description neighbor, with frozen order breaking ties. A second distribution weights all ordered nonself pairs by the original Gaussian kernel K_{ij}. We also report the small set with $K_{ij} \geq 0.5$.

Table 22: Observed overlap distributions. All overlap values and the $C \leq 25\%$ column are percentages. The latter is a fraction of directions or kernel mass. Weighted medians use the inverse empirical CDF, and other medians average the two central observations when needed.

Pair distribution	Directions	Mean C	Median C	$C \leq 25\%$	Median J
Nearest neighbor	148	39.2	36.1	44.6	20.1
Kernel-weighted pairs	21,756	33.5	20.3	56.9	10.0
$K \geq 0.5$ pairs	16	37.4	22.2	62.5	10.2

Nearest-neighbor directional coverage has median 36.1%, with 66 of 148 directions (44.6%) at or below 25%. The kernel-weighted mean is 33.5%. The half-height set comprises only 8 unordered

pairs across 13 benchmarks, with median directional coverage 22.2%. The nearest neighbor can be distant in semantic space, and shared columns make pair observations dependent. These distributions therefore receive no iid confidence intervals.

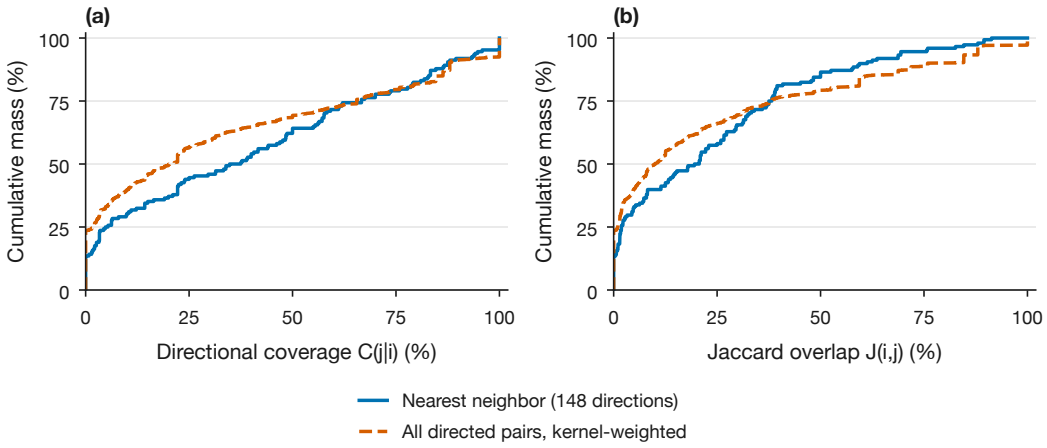

Figure 7: Low reporting overlap is present among description neighbors. Empirical distributions of (a) directional coverage and (b) Jaccard overlap use 148 nearest-neighbor directions or all 21,756 directed pairs weighted by semantic similarity. These are descriptive frozen-corpus distributions.

The observed low-overlap mass makes reporting-mask compatibility a practical scope condition. Its ranking consequences remain distinct from the partial-copy interventions: naturally neighboring columns differ in items and scores, and several reporting masks can coexist. Four identical copies sharing one mask provide a controlled mechanism test. Neither the overlap distribution nor those interventions estimate natural ranking-failure prevalence.

E.3 Metadata-restricted gains across evaluation families

We audit the saved density-equating outcomes without refitting. The broader support panel's retained target set has 112 paired seed–target outcomes: 20 targets contribute one seed and 46 contribute two. Averaging seeds within targets, then targets within families, reproduces the original mean +0.0241 across 66 targets and 43 families. Of those families, 28 have positive effects and 15 have negative effects, with median +0.0173. The original marginal interval and Holm $p = 0.6093$ across twenty comparisons remain the inferential result.

Table 23: Support through metadata restriction. The middle stage uses restricted-supported fitting rows while retaining all evaluable targets. The final stage also removes metadata-excluded targets.

Support	Evaluation scope	Seed–targets	Targets	Families
≥ 1	Original panel	231	120	60
≥ 1	Restricted rows, all targets	158	93	53
≥ 1	Restricted rows, retained targets	112	66	43
≥ 3	Original panel	81	50	29
≥ 3	Restricted rows, all targets	12	9	8
≥ 3	Restricted rows, retained targets	5	3	3

Table 24: Density-equating outcomes on identical restricted-supported rows and the same 66 retained targets. Entries average targets within each of 43 evaluation families, with equal family weights.

Training pools	Curated	Full	Full − curated
Original	0.7433	0.7941	+0.0508
Restricted	0.7908	0.8149	+0.0241

Both curated and full point estimates increase when the training pools are restricted on these fixed rows and targets. The curated estimate increases more, narrowing the retention gap. This four-cell comparison separates the training-pool contrast from target exclusion within this population. Changes in column selection, suite support and population limit a causal interpretation as protocol harmonization.

The strongest positive evaluation-family effect is LiveCodeBench (+0.3978), supported here by the retained `livecodebench-2408-2505` target. Removing that family from the aggregation gives +0.0152. Across all one-family deletions, the mean ranges from +0.0152 to +0.0281. Predictions remain fixed in this deletion analysis, so it is an influence summary, not a training refit or cross-validation. The five largest positive families contribute 49.8% of the sum of positive effects, with negative effects retained separately. Sparse family support precludes useful family-specific intervals in this audit. Figure 8 shows every surviving family and its support.

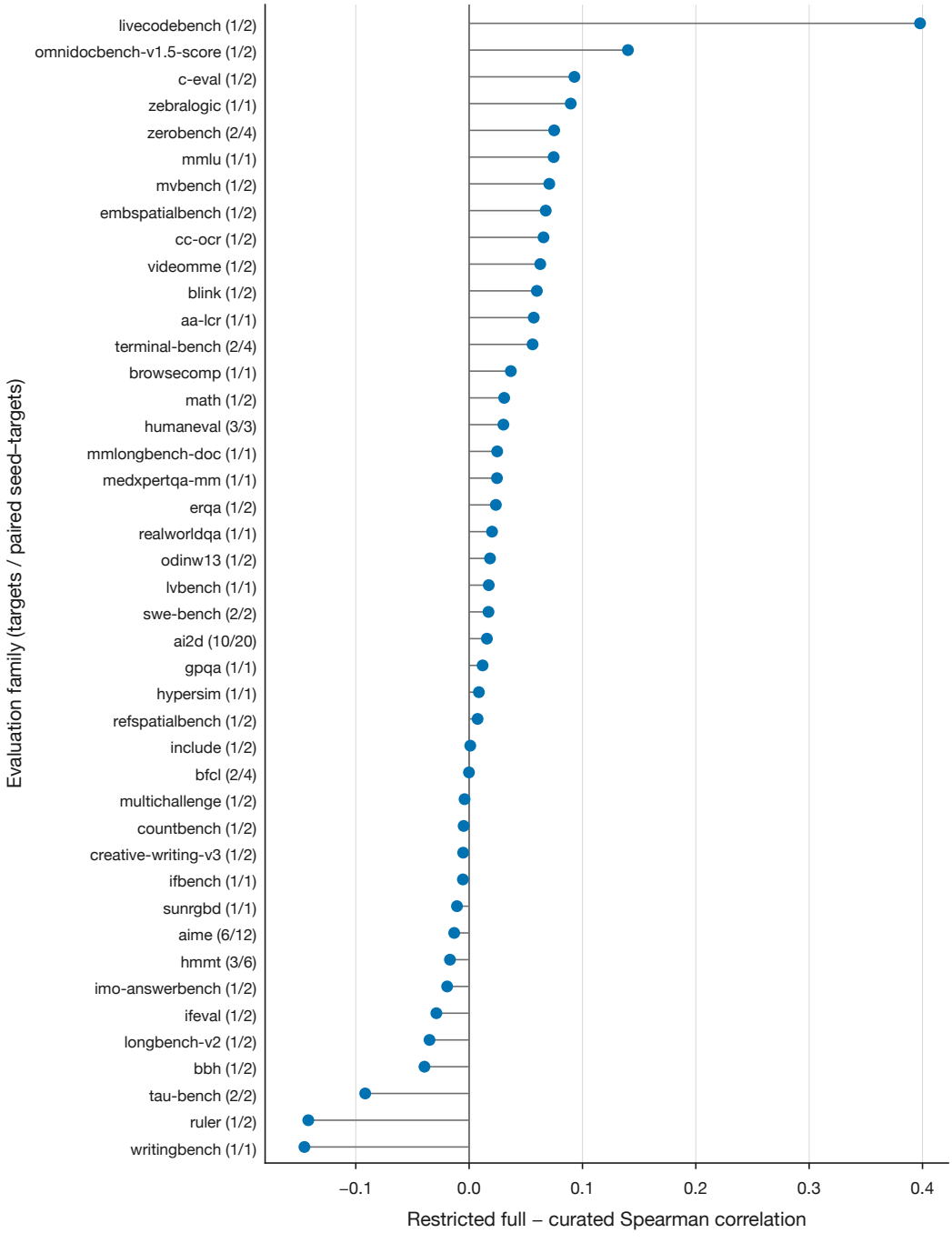

Figure 8: Metadata-restricted retention effects vary across families. Each point is the full-minus-curated effect after seed-within-target and target-within-family averaging. Parentheses give target and paired seed–target counts. All 43 surviving families are shown. No per-family confidence intervals or subgroup significance tests are added.

F Optional task-conditioned comparison

The extension uses global ability and nearby score residuals to predict performance on benchmark-like task queries. Its evaluation is separate from the fixed-population retention and re-listing analyses in the main text.

Task-conditioned model selection. Learned model and task representations support score prediction and routing (Zhuang et al., 2025; Song et al., 2025; Chen et al., 2025; Ge et al., 2026), and text-only forecasting predicts performance from task descriptions and model configurations (Park et al., 2025). Taxonomies, nearest neighbors, and hard neighborhoods define discrete relations among tasks, while kernels assign smoothly varying relevance (Li, 2025). BoB's optional extension uses semantic proximity to pool benchmark residuals that adjust global ability, and a support-matched top-k control separates neighborhood selection from correction strength.

F.1 Task-conditioned residual aggregation

The optional extension adjusts global ability using performance on benchmarks relevant to a query. Above-expected scores on nearby benchmarks give a positive adjustment. A query q is a task description embedded in the same space, and $g_{mi}(q) \geq 0$ specifies benchmark relevance. The prediction for model m combines shared ability and a residual correction:

$$\begin{aligned} \bar{r}_m^{(g)}(q) &= \frac{\sum_i g_{mi}(q)\bar{I}_i w_{mi} r_{mi}}{\sum_i g_{mi}(q)\bar{I}_i w_{mi}}, \\ P_m^{(g)}(q) &= \sum_i g_{mi}(q) w_{mi}, \\ y_m^{(g)}(q) &= \theta_m + \beta_m(q)\bar{r}_m^{(g)}(q). \end{aligned} \tag{5}$$

Unobserved cells have zero weight. The first line averages the model's residuals over relevant benchmarks, precision-weighted by curve information and the robust weights. The second line is the model's *local support* $P_m^{(g)}(q)$, the relevance-weighted amount of evidence it has near the query. The continuous rule uses $g_{mi}(q) = \exp[-d(q, x_i)^2/(2h^2)]u_i$, so relevance decays smoothly with distance and co-located entries share column mass at the chosen scale.

The coefficient $\beta_m(q)$ controls correction strength. BoB-Support uses $\beta_m(q) = \lambda P_m^{(g)}(q)/(P_m^{(g)}(q) + \alpha)$, with $\lambda = \alpha = 1$ in the original unit-scale rule. BoB-Constant uses $\beta_m(q) = c$. Both use the same residual mean and return θ_m at zero support. We treat support-dependent and constant coefficients as alternative regularization choices, comparing the unit-scale rule, inner-selected c, and joint selection of λ and α.

Controls. Relevance controls change only g, replacing it with hard category, nearest-neighbor, top-k, radius, cluster, or all-benchmark rules, or with a support-matched top-k rule (Section A.4). Global aggregation controls compare θ_m with uniform standardized averaging, a fixed curated subset, the first principal component (PC1), and score-covariance generalized least-squares (GLS) weighting (Section 4.1). These assess the complete aggregation procedures separately from the relevance controls.

F.2 Queries and score dependence

Prediction is evaluated on held-out benchmark descriptions. Each outer BoB fit excludes target scores. WildScores and the mechanism-matched AA controls select tunable task-relevance parameters on inner training folds. The original AA reference selected its relevance bandwidth across LOBO outcomes, and the fully nested control reaches the same reported level and profile values (Sections H.1.2 and H.6). Density weights depend on descriptions conditional on the retained list and density scale, while relevance-bandwidth selection uses training-fold outcomes. Query canonicalization is illustrated separately (Section H.7). Absolute-score calibration uses target scores from other release groups and is not a zero-shot prediction of an unseen benchmark's score scale.

F.3 RANKING AND SHORTLIST RESULTS

Q1 and Q2 test the two aggregation decisions. The extension asks whether global ability also supports task-conditioned comparison, on the original WildScores population. Eligibility is shared across methods within each metric (Table 25). Against the matched global BoB control at 0.817, BoB-Support gives a paired increment of +0.006 [+0.003, +0.011]. BoB-Constant exceeds unit-scale BoB-Support within the main grid (Holm $p = 0.006$), while the support-matched top-k contrast is nonsignificant (Holm $p = 0.486$). The result supports a regularized task-conditioned correction, with competitive relevance rules and correction strengths.

Table 25: Optional task correction on the original WildScores population. Ranking covers 133 targets in 67 families. Paired $\Delta\rho$ uses global θ as reference. MAE and Regret@3 (R@3) use 0–100 score points. Hit@3 (H@3) is a percentage. Eligibility is shared within each metric. MAE calibration withholds the evaluated release group.

Method	ρ	$\Delta\rho$ vs. θ	MAE	R@3	H@3 (%)
Uniform standardized averaging	0.764	-0.053	6.19	2.08	65.3
BoB global (θ only)	0.817	+0.000	5.18	1.62	71.7
BoB-Support	0.823	+0.006	5.10	1.67	72.6
BoB-Constant	0.830	+0.014	5.04	1.52	74.4

Shortlists select release-group representatives by training coverage. Regret measures the gap to the best observed candidate, and Hit records whether a best candidate is nominated (Section G).

BoB-Support reduces Regret@3 relative to uniform standardized averaging, from 2.08 to 1.67, and the paired reduction survives correction. It does not reduce regret relative to global BoB at 1.62 (paired change +0.055, Holm $p = 1.000$), separating ranking gains from shortlist utility. Section G reports all budgets, both correction variants, and paired effects.

Decision scope. Shortlists are retrospective and restricted to observed targets. Absolute-score calibration uses target observations from other release groups. Prospective task decisions, cost and latency constraints, and item-level analyses remain outside this validation.

Numerical scope. The canonical procedure uses six fitting rounds, with task bandwidth selected from five candidates on inner folds. Longer fits can reduce fixed-rule prediction, while repeated selection at 48 rounds or over stopping retains local ranking gains. Constant coefficients remain competitive. This finite-round procedure and its sensitivities are specified in Table 8 and Section G.1.

G WILDSCORES TASK CORRECTION AND SENSITIVITY

Tables below report the complete family-mean mechanism and shortlist results on WildScores. In these complete mechanism comparisons, rows below BoB-Support remove ($-$) its local correction or replace ($\rightarrow$) relevance g or the support factor. Each row changes the named part relative to BoB-Support. R@k is regret at budget k, and H@3 is best-candidate hit rate. Ranking, calibration, and shortlist eligibility differ. In particular, a target can support a shortlist comparison even when its score variation or configuration count makes ranking correlation undefined.

Table 26: WildScores, six-round fitting.

Method / change from Support	ρ	Profile	MAE	R@1	R@3	R@5	H@3 (%)
Uniform standardized averaging	0.764	0.084	6.19	4.79	2.08	1.66	65.3
BoB-Support	0.823	0.421	5.10	4.32	1.67	1.21	72.6
− local correction (θ only)	0.817	0.404	5.18	4.55	1.62	1.16	71.7
$g \to$ Category residual	0.757	0.343	6.34	5.36	2.19	1.37	64.5
$g \to$ Nearest residual	0.746	0.326	6.36	5.39	2.23	1.50	57.8
$g \to$ Nested top-k	0.764	0.327	6.35	5.42	2.01	1.30	61.5
$g \to$ Nested radius	0.817	0.400	5.20	4.51	1.62	1.16	71.3
$g \to$ Nested K-means	0.765	0.323	6.39	5.88	1.85	1.40	64.0
$g \to$ Nested agglomerative	0.765	0.335	6.32	5.50	1.73	1.27	65.7
$g \to$ Support-matched top-k	0.823	0.423	5.07	4.33	1.67	1.21	72.6
$g \to$ All-benchmark residual	0.768	0.284	6.17	5.15	1.92	1.22	66.4
BoB-Constant (factor $\to c$)	0.830	0.455	5.04	3.95	1.52	1.10	74.4

Table 27: WildScores, forty-eight-round fitting and repeated inner selection.

Method / change from Support	ρ	Profile	MAE	R@1	R@3	R@5	H@3 (%)
Uniform standardized averaging	0.764	0.084	6.19	4.79	2.08	1.66	65.3
BoB-Support	0.818	0.418	5.19	4.59	1.71	1.23	72.0
− local correction (θ only)	0.814	0.411	5.21	4.60	1.64	1.27	71.3
$g \to$ Category residual	0.745	0.336	6.42	6.11	2.16	1.54	65.2
$g \to$ Nearest residual	0.737	0.327	6.44	5.57	2.27	1.54	59.7
$g \to$ Nested top-k	0.747	0.313	6.50	5.57	2.20	1.46	60.6
$g \to$ Nested radius	0.812	0.412	5.23	4.60	1.64	1.27	70.9
$g \to$ Nested K-means	0.756	0.322	6.42	6.34	1.84	1.46	62.4
$g \to$ Nested agglomerative	0.755	0.325	6.40	5.70	1.74	1.46	64.7
$g \to$ Support-matched top-k	0.819	0.424	5.13	4.59	1.66	1.21	72.2
$g \to$ All-benchmark residual	0.718	0.263	6.52	5.95	2.07	1.34	61.4
BoB-Constant (factor $\to c$)	0.829	0.444	5.14	4.00	1.48	1.09	75.8

Table 28: WildScores paired effects at six rounds. Rank gains are positive, regret reductions negative. Intervals are paired family bootstrap intervals. Holm adjustment covers eleven contrasts per metric.

BoB-Support minus	$\Delta\rho$ [95% CI]	Holm p	ΔR@3 [95% CI]	Holm p
θ only	+0.006 [+0.003, +0.011]	0.001	+0.055 [-0.050, +0.227]	1.000
Uniform standardized averaging	+0.060 [+0.037, +0.083]	< 0.001	-0.404 [-0.708, -0.151]	0.014
Category residual	+0.066 [+0.044, +0.088]	< 0.001	-0.520 [-1.150, +0.009]	0.554
Nearest residual	+0.077 [+0.053, +0.102]	< 0.001	-0.563 [-1.065, -0.121]	0.208
Nested top-k	+0.059 [+0.038, +0.080]	< 0.001	-0.343 [-0.721, +0.033]	0.581
Nested radius	+0.006 [+0.003, +0.010]	0.001	+0.051 [-0.055, +0.223]	1.000
Nested K-means	+0.058 [+0.040, +0.077]	< 0.001	-0.181 [-0.461, +0.097]	1.000
Nested agglomerative	+0.058 [+0.040, +0.075]	< 0.001	-0.057 [-0.349, +0.236]	1.000
Support-matched top-k	+0.001 [-0.001, +0.002]	0.486	+0.003 [-0.043, +0.061]	1.000
All-benchmark residual	+0.055 [+0.040, +0.070]	< 0.001	-0.249 [-0.567, +0.052]	0.717
BoB-Constant	-0.007 [-0.012, -0.002]	0.006	+0.152 [+0.009, +0.347]	0.486

Table 29: WildScores domain summaries at six rounds. Means first aggregate within benchmark families. Small domains have limited coverage.

Domain	Rank targets	BoB-Support ρ	θ ρ	BoB-Support R@3	θ R@3
agentic	17	0.838	0.836	2.24	2.26
coding	15	0.865	0.845	0.45	0.46
conversation	1	0.784	0.800	5.05	5.05
instruction	4	0.826	0.817	0.77	0.74
knowledge	15	0.842	0.840	2.04	1.39
longcontext	8	0.713	0.711	1.52	1.52
math	13	0.830	0.794	0.28	0.26
multilingual	8	0.946	0.948	0.16	0.16
multimodal	44	0.829	0.823	2.03	2.02
other	3	0.772	0.762	0.97	1.36
reasoning	5	0.706	0.709	1.56	1.56

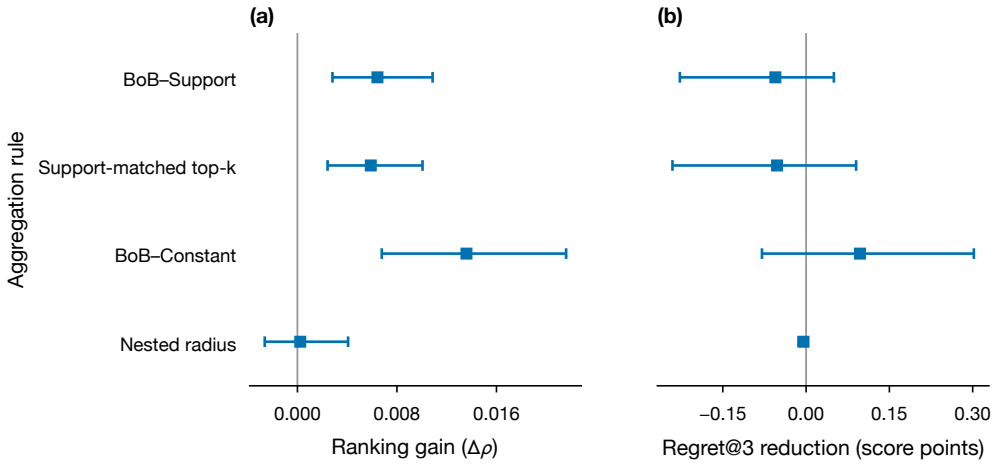

Figure 9: **Ranking and shortlist gains answer different questions.** Paired effects relative to θ on the original WildScores coverage. Positive values mean higher ranking correlation or lower three-candidate regret. Intervals are 5,000-resample family-bootstrap intervals conditional on the paired split seeds. Corrected tests are reported separately.

G.1 Sensitivity and numerical stability

The initial sensitivity protocol preceded its new runs. Follow-up extensions were added after observing stopping effects, boundary selections, and grid sensitivity. These result-motivated extensions provide exploratory evidence. The final reanalysis corrected source-lineage blocks and averaged tied initialization ranks, then repeated all dependent comparisons with unchanged scores, embeddings, seeds, and parameter grids. All results below use WildScores. Their paired intervals use unadjusted comparisons. The Holm corrections above apply to the subsequent mechanism comparisons.

Local correction scales. The initial nested grid crosses bandwidth multipliers $\{0.30, 0.45, 0.60\}$ with $\lambda \in \{0.25, 0.5, 1, 2\}$ and $\alpha \in \{0.25, 1, 4\}$, including θ. The constant search includes $c \in \{0.25, 0.5, 0.75, 1\}$. An expanded-grid follow-up lowers bandwidth to 0.15, strength to 0.125, and support scale to 0.0625. Both searches show useful local correction, with competitive constant coefficients. Selections can reach grid boundaries, locating preferred settings within the searched ranges. A further nested comparison selects stopping among 6,12,24,48 rounds jointly with local parameters and includes a stopping-tuned θ control.

Table 30: Local-scale searches at six rounds across two split seeds. The initial and expanded searches are both retained. MAE uses normalized 0–100 score points throughout the WildScores tables.

Setting	ρ	MAE
θ only	0.8167	5.178
Fixed unit BoB-Support	0.8210	5.472
Initial bandwidth selection	0.8246	5.084
Initial joint h, λ, α	0.8242	5.104
Initial constant selection	0.8256	5.133
Median-support coefficient	0.8256	5.203
No shrinkage	0.7784	6.118
Expanded joint search	0.8244	5.157
Expanded constant search	0.8303	5.038

Density and robustness. At fixed six-round fitting, $h = 0.45$, and unit local scales, the first split seed's five outer blocks vary density multipliers over $\{0.10, 0.15, 0.25, 0.40, 0.60\}$. Robust caps $\{2, 3, 4, 6\}$ are crossed with weight floors $\{10^{-2}, 10^{-3}, 10^{-4}\}$, alongside a no-robust-weight control. The tested floors produce identical metrics. Density affects prediction more substantially, with ranking and calibrated MAE favoring different settings.

Table 31: Density and robust-weight sensitivity on the first split seed. Robust rows use floor 0.001. The tested floors 0.01 and 0.0001 yield identical metrics. Each row changes one component of the fixed six-round rule.

Setting	ρ	MAE
Density multiplier 0.1	0.8078	5.569
Density multiplier 0.15	0.8117	5.573
Density multiplier 0.25	0.8248	5.406
Density multiplier 0.4	0.8234	5.105
Density multiplier 0.6	0.8206	5.113
Robust cap 2	0.8264	5.376
Robust cap 3	0.8248	5.406
Robust cap 4	0.8229	5.435
Robust cap 6	0.8234	5.445
No robust weighting	0.8234	5.445

The multiplicity study chooses eight benchmarks from distinct categories by coverage. Each perturbation adds four copies of its scores, using either identical or paraphrase vectors, and refits the full model including the distance median. Five density scales yield eighty perturbations. Mean absolute rank displacement is reduced relative to uniform standardized averaging, with nonzero worst-case displacement. Copied-score perturbations measure sensitivity to repeated listings. Fresh questions additionally contribute measurement information through new responses.

Table 32: Model-rank displacement after adding four copied-score listings. Means aggregate absolute displacement over models and eight perturbation targets. Worst movement is the largest individual change across those targets.

Density	Copies	Mean movement	Worst movement	Uniform mean
0.10	exact	0.152	6	5.907
0.10	paraphrase	0.764	48	5.907
0.15	exact	0.231	20	5.907
0.15	paraphrase	0.809	48	5.907
0.25	exact	0.626	44	5.907
0.25	paraphrase	0.955	55	5.907
0.40	exact	1.577	76	5.907
0.40	paraphrase	1.936	81	5.907
0.60	exact	1.859	69	5.907
0.60	paraphrase	2.092	70	5.907

Stopping and resolution. The instrumented fit exactly reproduces the canonical default arrays and field. Checkpoints compare 6,12,24,48 rounds, with a follow-up through 96 and 192 rounds. Late parameter updates shrink substantially, while fixed-rule held-out prediction deteriorates. This pattern motivates an early-stopping hypothesis. Repeated nested selection at 48 rounds preserves local ranking gains.

Table 33: Longer fitting and inner-selected stopping across two split seeds. These local-scale searches use the initial parameter grid.

Setting	ρ	MAE
48 rounds: θ	0.8139	5.214
48 rounds: fixed unit BoB-Support	0.8021	5.590
48 rounds: joint selection	0.8198	5.185
48 rounds: constant selection	0.8194	5.193
Stopping selection: θ	0.8150	5.208
Stopping + joint selection	0.8238	5.109
Stopping + constant selection	0.8233	5.153

At 48 rounds, grid steps 0.02, 0.01, and 0.005 give measurably different results. Separate tests expand the ability range from $[-3.5, 3.5)$ to $[-5, 5)$ and curve bounds from $a \in [0.2, 20], b \in [-4, 4]$ to $a \in [0.1, 40], b \in [-6, 6]$. The wider ability range has little effect, while some curve parameters remain on bounds. Per-curve optimizer status and across-round parameter changes track the two levels of optimization. The artifact retains the complete numerical grid and each round's changes, boundary hits and optimizer status.

Table 34: Selected sensitivity contrasts on the analysis corpus. Intervals use 5,000 paired family bootstrap resamples without multiplicity adjustment. Grid refinement uses the first split seed. The other rows use both seeds.

Paired contrast	$\Delta\rho$	95% family interval
Initial joint minus theta	+0.00745	[+0.00349, +0.01220]
Initial joint minus constant	-0.00139	[-0.00466, +0.00152]
Expanded joint minus constant	-0.00593	[-0.01000, -0.00198]
48-round joint minus theta	+0.00599	[+0.00184, +0.01092]
Stopping-selected joint minus theta	+0.00885	[+0.00471, +0.01396]
Step 0.005 minus step 0.01	-0.00133	[-0.00240, -0.00043]

Remaining fixed choices. Observation thresholds, family definitions, quadratic profile adjustment, residual-scale floors, and the source-conflict threshold remain fixed. These settings define the evaluation conditions. The study evaluates benchmark-like target descriptions using other scores from the observed model population.

H ARTIFICIAL ANALYSIS: DATA, CONTROLS AND ROBUSTNESS

H.1 DATA AND BASELINES

Artificial Analysis snapshot. We retrieved the snapshot directly from the Artificial Analysis Data API at `https://artificialanalysis.ai/api/v2/data/llms/models` at `2026-09-03T20:19:24Z`. The saved JSON contains 643 API rows and has SHA-256 digest `f2a230e9bd1fd55bfe058eaa59ae3f0401dbde76934609f4c5eef0adf57f745a`. We use AIME, AIME'25, MATH-500, GPQA Diamond, HLE, MMLU-Pro, LiveCodeBench, SciCode, IFBench, long-context reasoning, Terminal-Bench Hard, Terminal-Bench 2.1, τ^2-Bench, and τ-Bench Banking. A row enters the analysis when at least five of these fourteen evaluation fields are numeric, leaving 605 rows. Each API row is one observation, so separately listed model configurations remain separate observations. The population-shift analysis groups rows by the API's `model_creator.name` field.

A JSON `null` is represented as a missing cell and every numeric value, including zero, is retained as observed. The curve fit and residual field sum over observed cells. Robust cell weights attenuate large standardized residuals during fitting. Table 35 reports the resulting coverage.

Table 35: Cell coverage among the 605 retained Artificial Analysis API rows. Observed zero is a subset of observed. Null counts are values represented by JSON `null` in the saved snapshot.

Benchmark	Observed	Null	Observed zero
AIME	192	413	14
AIME'25	269	336	4
GPQA Diamond	604	1	0
HLE	602	3	0
IFBench	449	156	0
Long-context reasoning	535	70	64
LiveCodeBench	343	262	0
MATH-500	191	414	0
MMLU-Pro	345	260	0
SciCode	602	3	3
τ^2-Bench	439	166	23
τ-Bench Banking	203	402	0
Terminal-Bench Hard	432	173	48
Terminal-Bench 2.1	235	370	5

Table 36: Original AA target-prediction reference. Entries average fourteen target correlations, with standard errors in parentheses. The original relevance-bandwidth selection uses these fold outcomes. The fully nested mechanism-matched BoB control reaches the same level and profile values (Sections H.1.2 and H.6). The published AA Index cannot have its held-out target scores removed and is included only as a descriptive comparator.

Prediction method	Level ↑	Profile ↑
BoB-Support (ours)	**0.904** (0.014)	**0.483** (0.051)
Uniform standardized averaging	0.878 (0.016)	0.046 (0.042)
Raw-score category mean	0.864 (0.018)	0.303 (0.065)
Raw-score nearest benchmark	0.762 (0.027)	0.111 (0.037)
AA Intelligence Index	0.883 (0.019)	0.351 (0.095)
Methods fitted from the current score population		
Score-covariance weights	0.697 (0.058)	0.123 (0.068)
First principal component	0.823 (0.022)	0.026 (0.083)

H.1.1 BASELINE SPECIFICATION

Uniform standardized averaging averages the standardized pool scores. *Hand-assigned categories* averages the pool benchmarks sharing the held-out benchmark's label. The snapshot provides no tax-

onomy, so the labels are ours: mathematics (AIME, AIME'25, MATH-500), knowledge (GPQA Diamond, HLE, MMLU-Pro), coding (LiveCodeBench, SciCode), agentic (τ^2-Bench, τ-Bench Banking, and both Terminal-Bench variants), instruction following (IFBench), and long context. The last two categories contain one benchmark each. When that benchmark is held out, the category rule falls back to equal weighting. *Nearest benchmark* returns the standardized score of the closest observed pool benchmark in description space.

The *Artificial Analysis Intelligence Index* is used exactly as published. The API supplies composite values without a version identifier or formula. The ten-column list used for our descriptive split corresponds to the historical v3 suite. AA's version history records subsequent component changes (Artificial Analysis, 2026b), so this list does not establish overlap with the composite returned in September 2026. Among the 248 models observed on all ten historical columns, the supplied index correlates at Spearman 0.938 with their equal-weight raw mean. We use the supplied values directly as a descriptive comparator. It cannot be refitted with targets withheld.

Two score-derived aggregation controls estimate structure from the current observed score population. *Score-covariance GLS* uses $w \propto \Sigma^{-1}\mathbf{1}$, where Σ is the empirical correlation matrix of row-centered standardized pool scores with ridge 0.05. *First principal component* takes the leading component of the standardized pool matrix after filling missing cells with column means and aligns its sign with the mean score. Both procedures serve as score-derived aggregation controls.

H.1.2 Mechanism-matched aggregation controls

All aggregation controls instantiate Equation (5) with the same outer-fold characteristic curves, θ_m, standardized residuals, mean item information, robust cell weights, support regularizer, and missing-cell treatment. BoB uses $g_{mi}(q) = \exp(-d(q, x_i)^2/(2h^2))u_i$.

The category-residual rule sets $g_{mi}(q) = 1$ when pool benchmark i shares the query category and zero otherwise. IFBench and long-context reasoning are singleton categories, so their held-out folds have zero category support and return θ_m. The nearest-residual rule gives unit relevance to each model's closest observed pool benchmark. Top-k selects its k closest observed pool benchmarks. Radius relevance includes every pool benchmark within its threshold. K-means and agglomerative rules fit clusters on the pool descriptions and assign the query to the closest normalized cluster centroid. The agglomerative rule uses cosine distance and average linkage. K-means uses 30 initializations with seed zero.

The support-matched top-k rule begins with a model's observed continuous weights a_{mi}. Its neighbor count is the rounded Kish quantity $(\sum_i a_{mi})^2/\sum_i a_{mi}^2$. It selects that many nearest observed benchmarks and rescales their common hard weight so that $\sum_i g_{mi}(q)w_{mi}$ exactly equals the continuous support mass. Across all outer folds, the maximum numerical discrepancy is below 5×10^{-16}.

Each outer fold computes its distance unit from the median pairwise distance among the thirteen pool descriptions. Inner folds select h from $\{0.30, 0.45, 0.60\}$ times that distance, k from $\{1, 2, 3, 4, 5\}$, radius from $\{0.25, 0.40, 0.55, 0.70, 1.00\}$ times that distance, and cluster count from $\{2, \ldots, 8\}$. Mean inner held-out profile is the selection criterion. The outer benchmark supplies the prediction target after every fit, scale, and hyperparameter has been fixed.

H.2 Additional Artificial Analysis illustrations and controls

The AA controls below use BoB-Support as the reference aggregation variant. Figure 10 shows how neighborhood density in the retained AA description set determines individual semantic weights. The score-equating illustration in Figure 2 uses the same frozen snapshot. Its residual example is DeepSeek R1 Distill Llama 70B on AIME.

H.2.1 Ablations and controls

Characteristic curves alone reach 0.194 profile and a residual field on the linear scale reaches 0.100, versus 0.483 for the support-dependent variant (Table 38). These constructions use their own ability controls. Removing shrinkage from BoB-Support lowers level and profile to 0.845 and 0.392. A constant coefficient gives 0.483 profile and 0.907 level. This AA control fixes the coefficient to the median support factor across models in each held-out fold. The release study selects the coefficient

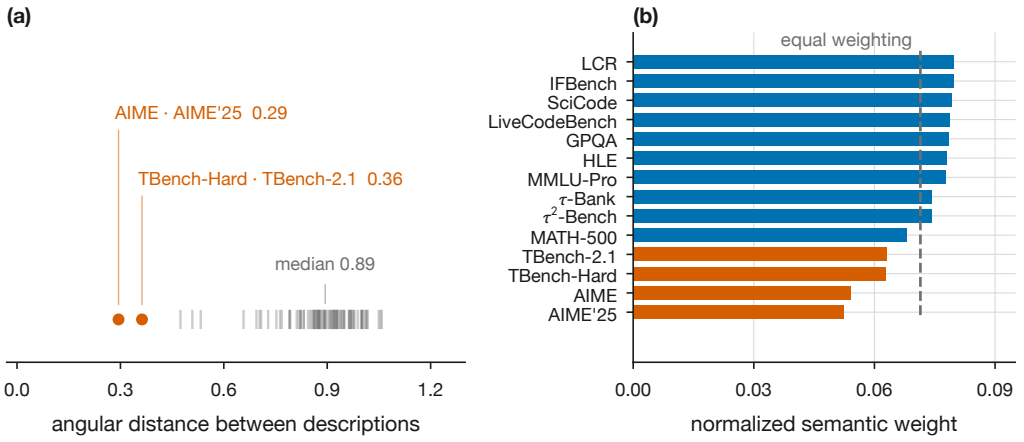

Figure 10: **Description-set density determines semantic benchmark weights. (a)** The two closely related benchmark pairs lie far below the median of all 91 angular distances among 14 descriptions. **(b)** Inverse-density weighting lowers their individual influence below the $1/14$ share assigned by equal weighting.

Table 37: Held-out profile under mechanism-matched aggregation rules. All rows share the fitted score-side components. Relative to BoB-Support, $g \to$ replaces relevance and $-$ removes the local correction. The final column gives one-sided Holm adjustment over seven discrete controls for quadratic profile.

Method / change from Support	Profile ↑	Isotonic ↑	p_{Holm}
BoB-Support	**0.483** (0.051)	**0.464** (0.052)	–
$g \to$ Support-matched top-k	0.454 (0.054)	0.443 (0.056)	0.104
$g \to$ Category residual	0.410 (0.055)	0.404 (0.055)	0.104
$g \to$ Nested top-k	0.388 (0.058)	0.386 (0.059)	0.010
$g \to$ Nearest residual	0.357 (0.053)	0.354 (0.051)	0.010
$g \to$ Nested radius	0.351 (0.049)	0.346 (0.051)	0.004
$g \to$ Nested K-means	0.281 (0.060)	0.275 (0.060)	< 0.001
$g \to$ Nested agglomerative	0.293 (0.058)	0.292 (0.058)	< 0.001
$-$ local correction (θ only)	0.397 (0.045)	0.380 (0.048)	–

through inner validation. Both regularization choices preserve the predictive value of the equated residual field.

The geometry controls separate prediction from multiplicity correction. Setting $u_i \equiv 1$ gives 0.473 profile while increasing ability-scale copy shift from 0.095 to 0.963 standard deviations. Reassigning descriptions and rerunning the method 200 times gives mean profile 0.358 and maximum 0.447, below 0.483 in every draw. This null differs from the 0.357 nearest-residual control, which preserves geometry and uses one observed neighbor. Paraphrased and lineage-stripped descriptions retain the result, and three embedding sizes move level and profile by at most 0.002 (Sections H.5 and H.5.1).

H.2.2 Empirical stability under population shift

Adding 43–81 models from one developer leaves the pre-existing BoB ranking at approximately 0.9999 or higher (Table 39). Score-covariance weights move it to between 0.9930 and 0.9986. This experiment measures downstream population stability of the full fitted ranking. The description-derived semantic weights remain fixed while score-side quantities are refitted.

Table 38: Standalone constructions and matched controls on AA. Rows A to D use construction-specific ability controls and compare complete procedures, not isolated component effects. Rows below BoB-Support share its evaluation path and remove (−) or replace (→) the named component. Presence, absence, and constant shrinkage are marked by +, −, and const. Ability shift is the mean scenario-wise maximum change in standardized ability after four added copies. Dashes indicate unmeasured controls.

Construction / change	Curves	Field	Shrink.	Level ↑	Profile ↑	Ability shift ↓
A: Linear z, precision mean	−	−	−	0.891	0.041	0.754
B: Characteristic curves	+	−	−	0.898	0.194	0.682
C: Linear z field	−	+	−	0.896	0.100	0.337
D: Curves + residual field	+	+	−	0.837	0.296	1.912
BoB-Support	+	+	+	**0.904**	**0.483**	**0.095**
− shrinkage	+	+	−	0.845	0.392	–
BoB-Constant (factor → c)	+	+	const.	0.907	0.483	–
− density weighting ($u_i \equiv 1$)	+	+	+	0.899	0.473	0.963

Table 39: Population-shift test. Each row adds models from one developer. Entries are Spearman correlations between the pre-existing models' rankings before and after the addition.

Developer added	Models added	Score-covariance weights	BoB-Support
Google	61	0.99433	**0.99995**
OpenAI	81	0.99298	**0.99988**
Anthropic	43	0.99859	**0.99991**

H.3 Held-out evaluation details

Table 40: Held-out performance split by membership in the historical AA v3 suite. This is not a verified overlap split for the supplied Index. Bold indicates the best entry in each column.

	Historical v3 suite (10 folds)		outside it (4 folds)	
Prediction method	Level ↑	Profile ↑	Level ↑	Profile ↑
BoB-Support (ours)	**0.890**	**0.424**	**0.939**	**0.632**
Uniform standardized averaging	0.874	0.076	0.887	−0.030
Hand-assigned category	0.857	0.283	0.882	0.352
Nearest benchmark	0.764	0.119	0.758	0.089
AA Intelligence Index	0.882	0.343	0.887	0.371

The historical suite contains 10 of the 14 benchmark columns (Table 40). On the four other targets, BoB reaches 0.939 level and 0.632 profile, versus 0.887 and 0.371 for the supplied index. This split describes performance across historical suite membership. It does not isolate the effect of target overlap in the published composite.

What level and profile measure. Level correlation largely reflects general ability. Quadratic profile removes its fitted trend in ranks, while residual monotone dependence can remain. The isotonic profile below controls this broader class of transformations. The 10-fold close-pair sensitivity excludes AIME/AIME'25 and Terminal-Bench Hard/2.1.

Table 41: Paired profile comparisons between BoB and each baseline. Entries are mean per-fold differences, one-sided Wilcoxon signed-rank p-values, and Holm-adjusted p-values.

	All 14 folds			10 folds excluding the two primary close-pair families		
BoB-Support minus	Δ profile	p	p_{Holm}	Δ profile	p	p_{Holm}
Uniform standardized averaging	+0.438	< 0.001	< 0.001	+0.367	< 0.001	0.006
Hand-assigned category	+0.181	0.008	0.017	+0.184	0.032	0.073
Nearest benchmark	+0.373	< 0.001	< 0.001	+0.339	< 0.001	0.006
AA Intelligence Index	+0.133	0.292	0.292	+0.072	0.423	0.423
Score-covariance weights	+0.361	0.003	0.010	+0.265	0.024	0.073
First principal component	+0.457	0.002	0.008	+0.367	0.010	0.039

For Table 42, the binned control is the sample-size-weighted mean Spearman correlation within ability deciles. Bins with fewer than 15 models or a degenerate variable are omitted.

Table 42: Profile correlation after three controls for general ability. The final column reports one-sided Wilcoxon tests against BoB under isotonic residualization.

Prediction method	Isotonic	Binned	Quadratic	p (isotonic)
BoB-Support (ours)	**0.464**	**0.498**	**0.483**	–
Uniform standardized averaging	0.003	0.211	0.046	< 0.001
Hand-assigned category	0.283	0.298	0.303	0.008
Nearest benchmark	0.106	0.181	0.111	< 0.001
AA Intelligence Index	0.337	0.387	0.351	0.313
Pool ability alone	0.001	0.253	0.097	< 0.001
Calibrated θ alone	0.381	0.431	0.396	0.003

Under isotonic residualization, θ alone reaches 0.381 and BoB reaches 0.464. Their paired difference is significant ($p = 0.0026$), locating the task-conditioned increment beyond score equating. The paired tests over all 14 folds and the primary close-pair sensitivity are reported in Table 41. In the 10-fold sensitivity, equal weighting, nearest benchmark, and first principal component survive Holm correction.

H.3.1 AGGREGATION-CONTROL RESULTS

Table 43 reports the paired profile effects of every aggregation control.

Table 43: Paired BoB-minus-control profile effects. Intervals are percentile 95% bootstrap intervals over 14 benchmark folds. p_W, p_{SF}, and p_H denote one-sided Wilcoxon, exact sign-flip, and Holm-adjusted Wilcoxon values.

Metric	Control	Δ	95% interval	Wins	p_W	p_{SF}	p_H
Profile	Support-matched top-k	+0.029	[-0.001, +0.060]	10/14	0.052	0.049	0.104
	Category residual	+0.073	[-0.018, +0.175]	9/14	0.077	0.091	0.104
	Nested top-k	+0.095	[+0.049, +0.141]	11/14	0.003	0.002	0.010
	Nearest residual	+0.126	[+0.056, +0.211]	10/14	0.003	< 0.001	0.010
	Nested radius	+0.132	[+0.073, +0.192]	12/14	< 0.001	< 0.001	0.004
	Nested K-means	+0.202	[+0.142, +0.264]	13/14	< 0.001	< 0.001	< 0.001
	Nested agglomerative	+0.190	[+0.133, +0.260]	14/14	< 0.001	< 0.001	< 0.001
Isotonic	Support-matched top-k	+0.021	[-0.009, +0.053]	8/14	0.163	0.111	0.268
	Category residual	+0.060	[-0.032, +0.160]	9/14	0.134	0.136	0.268
	Nested top-k	+0.077	[+0.030, +0.124]	9/14	0.010	0.006	0.030
	Nearest residual	+0.110	[+0.043, +0.191]	10/14	0.003	0.002	0.010
	Nested radius	+0.118	[+0.062, +0.176]	12/14	< 0.001	0.001	0.004
	Nested K-means	+0.188	[+0.129, +0.248]	13/14	< 0.001	< 0.001	< 0.001
	Nested agglomerative	+0.172	[+0.113, +0.241]	14/14	< 0.001	< 0.001	< 0.001

All 14 outer folds select continuous bandwidth 0.45. Top-k selects two or three neighbors. Radius and cluster count vary by fold, so their reported scores use the selection made without the outer target.

Per-fold predictions and selected parameters are included in the artifact. Excluding the two singleton-category folds leaves 12 comparisons. The mean BoB-minus-category difference is 0.085 for quadratic profile and 0.071 for isotonic profile. Their one-sided Wilcoxon values are 0.076 and 0.117. The positive average category contrast therefore persists when every held-out category has at least one pool peer.

Table 44: Profile under a 19-model population addition. The 586-model snapshot uses the same descriptions and nested procedure as the 605-model snapshot.

Aggregation rule	586 models	605 models	Change
BoB-Support continuous kernel	0.462	0.483	+0.021
Support-matched top-k	0.438	0.454	+0.016
Category residual	0.404	0.410	+0.006
Nested top-k	0.384	0.388	+0.004
Nearest residual	0.347	0.357	+0.010

The earlier snapshot's 586 retained rows all occur in the later snapshot, which adds 19 rows. All 5,327 benchmark cells observed in both snapshots are exactly unchanged. Table 44 therefore measures stability under a population addition, a distinct estimand from benchmark-outcome replication.

H.4 Ablations and validity checks

Rows A–D of Table 38 are self-contained constructions. They use their construction-specific ability control, whereas the full-method controls use the pool-only evaluation path of the headline experiment. The rows compare complete constructions with interacting design choices. The additional linear z-score and precision-weighted-mean construction reaches 0.041 profile.

H.4.1 Geometry controls

Flat weighting. Setting $u_i \equiv 1$ throughout the otherwise unchanged method gives 0.899 level, 0.473 profile, and an ability shift of 0.963 standard deviations, versus 0.095 for BoB. The one-sided profile comparison does not resolve a gain from density weighting ($p = 0.121$). Inverse-density weighting contributes multiplicity robustness.

Shuffled descriptions. Reassigning descriptions to benchmarks and rerunning the complete leave-one-benchmark-out procedure 200 times yields profile 0.358 ± 0.042 (maximum 0.447), versus the observed 0.483. This separates the contribution of correct semantic placement from generic pooling of calibrated residuals.

Residual-to-ability scale. Equation (5) adds an average standardized residual to standardized ability. A first-order latent-offset alternative divides each cell residual by the local derivative of its characteristic curve and weights by local information. It gives 0.907 level and 0.459 profile, providing a sensitivity check on the reported direct standardized-residual convention.

Support-factor specification. Including normalized item information inside $P_m^{(g)}(q)$ gives 0.907 level and 0.484 profile. The two-sided comparison does not detect a profile difference ($p = 1.00$) from including information in both the residual average and its support factor.

H.5 Text and representation robustness

Paraphrased descriptions. Replacing every benchmark description with an independently generated paraphrase gives 0.905 level and 0.493 profile, preserving the held-out result.

Lineage-stripped descriptions. Four generated descriptions explicitly name a relationship to another benchmark. We remove every cross-benchmark reference, re-embed, and rerun the analysis. AIME and AIME'25 move from distance 0.295 to 0.227 and remain the closest pair. The Terminal-Bench variants move from 0.363 to 0.461 and remain second. Held-out prediction remains 0.903 level and 0.475 profile, compared with 0.904 and 0.483. The τ pair, τ^2-Bench and τ-Bench Banking, shares conversational tool use, simulated users and policy constraints across different domains. Its distance is 0.509 (0.647 after stripping), well above the AIME pair and nearer the list median of 0.89. Density weighting therefore treats the two as largely distinct despite their shared protocol features.

H.5.1 Embedding size

Table 45: Primary quantities recomputed with three sizes of the Qwen3 embedding family (Zhang et al., 2025). The main text uses the 8B column.

	Qwen3-Emb 0.6B	Qwen3-Emb 4B	Qwen3-Emb 8B
AA level	0.903	0.904	0.904
AA profile	0.483	0.482	0.483
AA maximum rank displacement (equal weight 98.9)	24.1	15.8	15.9

Level and profile vary by at most 0.002 across sizes (Table 45). Four-copy displacement is 24.1 ranks for 0.6B, 15.8 for 4B, and 15.9 for 8B, below equal weighting's 98.9.

H.6 Sensitivity and selection

Field bandwidth. The field bandwidth is selected over $\{0.3, 0.45, 0.6\}$ times the median pairwise distance, using the mean of level and profile under leave-one-benchmark-out validation. The criterion is 0.669, 0.694, and 0.668, selecting $h = 0.402$. A leave-one-fold selection sensitivity uses the other thirteen fold outcomes to choose h. Every fold selects 0.45 times the median, giving 0.904 level and 0.483 profile. Those thirteen fits can include the outer target in their training pools, so this sensitivity is not fully nested validation. The mechanism-matched comparison uses the outer-pool distance scale defined in Section H.1.2. Every fold again selects 0.45 times its pool median, giving 0.904 level and 0.483 profile.

Density scale. The rule $h_{\mathrm{dens}} = 0.25$ times the median distance gives $h_{\mathrm{dens}} = 0.224$ and $W = 12.5$. Across Figure 11a, W falls from 14.0 at 0.05 times the median distance to 1.5 at the median itself. Effective benchmark mass is indexed by this explicit scale.

Robust constants and residual scales. Sweeping the robust cap over $\{2, 3, 4, 6\}$ and its floor over $\{10^{-2}, 10^{-3}, 10^{-4}\}$ moves level by at most 0.004 and profile by at most 0.007. The robust weight floor is inert over this range. For each benchmark, σ_i is the population standard deviation of its observed fitted residuals, with a numerical lower bound of 10^{-4}. The information calculation uses $\max(\sigma_i^2, 10^{-6})$ in its denominator. These numerical stabilizers bound residual standardization and information weights.

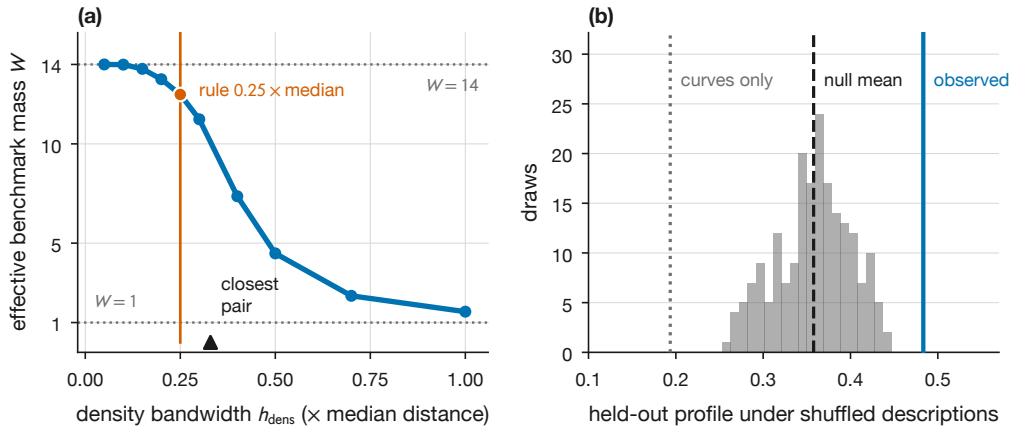

Figure 11: **Density-scale sensitivity and shuffled-description controls. (a)** Effective mass W across density scales, with the 0.25-times-median rule and closest observed pair marked. **(b)** Complete-method profile across 200 shuffled description runs. The null mean is 0.358 and no draw reaches the observed 0.483.

H.7 Query canonicalization

Table 46: Effect of query canonicalization. The support coefficient is the median over models of $\beta_m(q) = P_m^{(g)}(q)/(P_m^{(g)}(q) + 1)$.

	Distance to nearest benchmark ↓		Median support factor	
Task query	One sentence	Canonical	One sentence	Canonical
customer support automation	1.01	0.81	0.18	0.31
long report analysis	0.92	0.75	0.25	0.34
math competition coach	0.70	0.59	0.22	0.32
poetry writing partner	1.06	1.00	0.12	0.21
software engineering agent	0.99	0.70	0.20	0.39
Median benchmark–benchmark distance		0.89		

Benchmark descriptions share a template that names the measured subject, format, and difficulty. Canonicalization rewrites a user's query into that register before embedding. The four queries with a corresponding benchmark family move closer to the list, while the poetry query remains the farthest. The resulting support coefficient (Table 46) follows that geometric proximity and sets the regularization strength of the query adjustment.

H.8 Illustrative software-agent query

Consider the query "Act as an autonomous software engineering agent that works in a terminal: plan multi-step changes, run build and test commands, edit files, and debug failures in a real repository without human intervention." Canonicalization yields a description of autonomous software engineering in a real repository that names planning, editing, testing, debugging, tool use, code comprehension, and recovery across long sessions. Its nearest benchmarks are Terminal-Bench Hard at distance 0.701, Terminal-Bench 2.1 at 0.805, and LiveCodeBench at 0.876. Median support is 0.387. The largest absolute rank movers are `devstral-2` from 402 to 281, `gemini-3-5-flash-minimal` from 197 to 80, and `north-mini-code` from 278 to 167. The resulting top three are `claude-fable-5-1`, `claude-fable-5-1-xhigh`, and `claude-fable-5-1-high`. This example illustrates the computation. Held-out benchmarks provide the quantitative validation.

H.9 GLOBAL-RANKING COMPARISON

Across the 603 model configurations with both scores, the two full rankings agree at Spearman $\rho = 0.970$ and Kendall $\tau = 0.862$. The *Observed* column counts non-missing scores among the fourteen analyzed benchmarks.

Table 47: Union of the top 20 model configurations under the global BoB ability estimate θ_m and the AA Intelligence Index. Bold indicates top-20 membership. This is a descriptive comparison of complete rankings, not a target-withheld prediction test.

Model configuration	Benchmarks observed	BoB rank	θ_m	AA rank	AA Index
Claude Fable 5.1 (Adaptive Reasoning, Max Effort, Default Fallback)	6	**1**	1.778	**1**	65.7
Claude Fable 5.1 (Adaptive Reasoning, Xhigh Effort, Default Fallback)	6	**2**	1.739	**2**	64.8
Muse Spark 1.3 (max)	6	**3**	1.685	**6**	62.1
Claude Fable 5.1 (Adaptive Reasoning, High Effort, Default Fallback)	6	**4**	1.659	**4**	62.5
Claude Opus 5 (Adaptive Reasoning, Xhigh Effort)	6	**5**	1.630	**4**	62.5
Claude Opus 5 (Adaptive Reasoning, Max Effort)	6	**6**	1.627	**3**	63.1
Claude Opus 5 (Adaptive Reasoning, High Effort)	6	**7**	1.625	**8**	61.5
Muse Spark 1.3 (xhigh)	6	**8**	1.623	**13**	60.8
Kimi K3 (max)	6	**9**	1.620	**17**	59.7
Gemini 3.8 Flash (high)	6	**10**	1.609	22	58.7
GPT-6 Astra (xhigh)	6	**11**	1.609	**10**	61.0
GPT-6 Astra (max)	6	**12**	1.608	**9**	61.2
GPT-5.6 Sol (max)	9	**13**	1.607	**11**	60.9
Claude Fable 5.1 (Adaptive Reasoning, Medium Effort, Default Fallback)	6	**14**	1.605	**14**	60.5
Grok 4.6 (high)	6	**15**	1.596	**11**	60.9
GPT-6 Astra (high)	6	**16**	1.587	**15**	60.3
Claude Fable 5 (Adaptive Reasoning, Max Effort, Opus 4.8 Fallback)	9	**17**	1.570	**6**	62.1
GLM-5.3 (max)	6	**18**	1.570	**18**	59.5
Gemini 3.8 Flash (medium)	6	**19**	1.551	32	56.6
Qwen3.8 Max	6	**20**	1.545	24	58.1
GPT-6 Astra (medium)	6	22	1.539	**19**	59.2
Grok 4.6 (xhigh)	6	24	1.529	**16**	60.0
GPT-5.6 Sol (xhigh)	9	27	1.520	**20**	59.0
Grok 4.6 (medium)	6	30	1.493	**20**	59.0