POPULATION-CONDITIONED BENCHMARK ITEM INFLUENCE BEYOND DIFFICULTY AND DISCRIMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark item importance is often summarized by pass rate and discrimination. For model ranking, however, these summaries leave a population-conditioned component: **which items matter depends on which models are being compared**. We introduce **population-conditioned ranking influence**, a close-pair leverage statistic measuring which items distinguish similarly performing models within a specified population. Across nine benchmark matrices, residual fingerprints retain reproducible population-specific structure after controlling for population-local pass rate, pass rate squared, and corrected item–total discrimination, including under target-excluded capability definitions. On an exactly aligned 441-model by 12,032-item MMLU-Pro matrix, **fully disjoint panels independently recover this structure**: increasing panel size from 75 to 100 raises median within-population Spearman similarity from 0.361 to 0.435, while cross-population similarity remains much smaller, changing from 0.072 to 0.081. Fixed-margin Curveball randomization does not reproduce this agreement. The same disjoint-panel ordering appears across four additional large response matrices spanning scientific questions, multi-step reasoning, instruction following, and hard mathematics, with GPQA providing a second strong fixed-margin-null-confirmed result. Finally, signed item sets learned from one capability population preserve close-pair rankings of unseen models from that population better than sets learned from mismatched populations; all six paired contrasts are positive with 95% confidence intervals excluding zero. Together, these results show that **benchmark ranking importance is relational rather than intrinsic**: which questions matter depends on which models are being distinguished, and that population-specific structure predicts relative item-set utility on unseen models.

1 INTRODUCTION

Benchmark evaluations are usually summarized by aggregate scores, while individual questions are characterized by difficulty and discrimination. When two models achieve nearly the same overall score, a different question becomes important: **which items actually determine their ordering?** The same item can distinguish models in one capability regime while contributing little to ranking another. Its ranking importance therefore depends on the evaluated model population.

We formalize this dependence with *population-conditioned ranking influence*, $I(q \mid P)$, which measures how often item q separates close-scoring model pairs within population P. Population dependence alone would be unsurprising: items become easier or harder as capability changes, and their discrimination changes too. We therefore recompute population-local pass rate and corrected item–total discrimination, include a quadratic pass-rate term, and regress these effects out of influence. The resulting *residual influence fingerprint* captures ranking structure unexplained by these specified controls.

We organize the evidence around three questions. *Does the phenomenon appear broadly?* Across nine matrices spanning mathematics, knowledge, medicine, law, commonsense reasoning, social bias, reading comprehension, factual recall, and multilingual examinations, fingerprints from 90% model subsamples closely resemble their full-population references. Median Spearman correlations

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

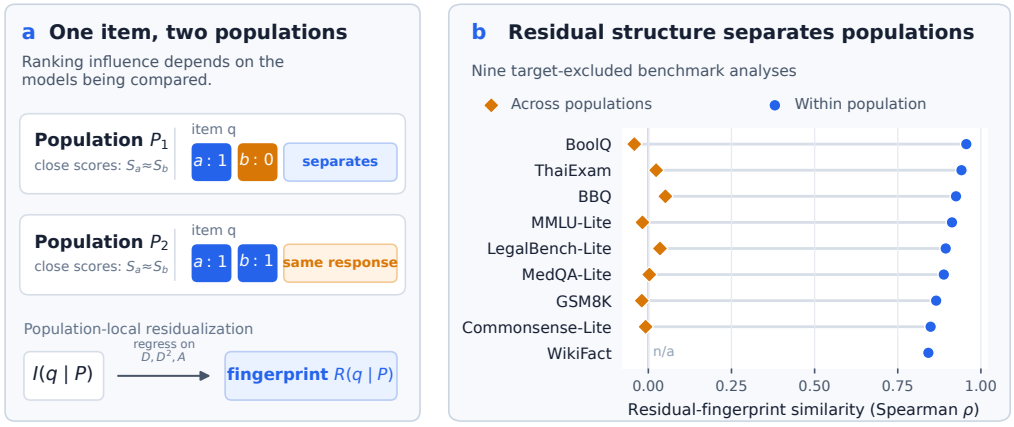

Figure 1: **Population-conditioned influence and breadth result.** Left: the same item can separate a close pair in one population but not another; influence is residualized against population-local pass rate and corrected discrimination. Right: across nine benchmarks, 90% subsample fingerprints closely match full-population references, whereas cross-population fingerprints have near-zero similarity. WikiFact has no analyzable cross-population pair.

range from 0.842 to 0.956, while analyzable cross-population correlations range from -0.042 to 0.051. Capability groups exclude the target benchmark, preventing direct reuse of its responses in population assignment.

Can the signature be recovered independently? A $441 \times 12{,}032$ MMLU-Pro matrix supports fully disjoint panels. We then repeat that design on MuSR, IFEval, MATH Level 5, and GPQA, each with more than 400 aligned models. Same-population fingerprints agree substantially more than cross-population fingerprints across these distinct target tasks, and fixed-margin Curveball randomization does not reproduce the agreement.

Does the signature predict what matters for new models? Signed item sets learned from low-, middle-, and high-capability selection panels each perform best, among the three selectors, on held-out models from the matching regime. All six matched-versus-mismatched effects are positive with intervals excluding zero.

Our contributions are:

1. **Population-conditioned ranking influence.** A close-pair item statistic and residual fingerprint beyond specified population-local item controls.
2. **Structural evidence across domains and at scale.** Nine-benchmark breadth, five disjoint-panel target matrices, panel-support diagnostics, and fixed-margin nulls.
3. **Predictive transfer.** Signed item sets preserve close rankings preferentially for unseen models from the capability regime in which they were learned.

2 Related Work

Psychometric evaluation. Item response theory estimates latent ability alongside item difficulty and discrimination. It provides a natural framework for language-model evaluation, although many items, few models, and clustered capabilities can make item estimates unstable (Jiang et al., 2026). We ask whether conventional local item statistics explain which questions resolve close model rankings, and use them as controls.

Capability-adaptive benchmarking. Fluid Benchmarking uses IRT to select questions near a model's inferred ability (Hofmann et al., 2025). We share its premise that item usefulness can depend on capability. Our contribution concerns reproducible *residual ranking influence*: population-

specific structure in the items that distinguish closely performing models after controlling for local pass rate and discrimination.

Efficient ranking evaluation. Micro-benchmarking studies small subsets that preserve model rankings, including comparisons between closely scoring models (Yauney et al., 2026); MetaEval identifies discriminative items for efficient score and ranking approximation (Wang et al., 2026). Our held-out experiment connects item analysis with ranking evaluation by testing whether the population used to select items predicts their relative utility for unseen models.

Benchmark diagnostics. Fantastic Bugs uses response patterns to identify potentially problematic items (Truong et al., 2025). Population-specific influence need not indicate an anomaly: a valid question can distinguish one population and be uninformative for another. We study this ranking-specific relationship between items and model populations.

3 Population-Conditioned Ranking Influence

3.1 Close-pair ranking influence

Let $Y \in \{0,1\}^{M\times Q}$ be a binary response matrix, with $Y_{aq} = 1$ when model a answers item q correctly and full-benchmark score $S_a = \sum_{q=1}^{Q} Y_{aq}$. For population P and score margin δ, define

$$C(P,\delta) = \{(a,b) : a,b \in P,\ a < b,\ |S_a - S_b| \le \delta\}.$$

Item q's close-pair leverage and normalized influence are

$$L(q \mid P,\delta) = \sum_{(a,b)\in C(P,\delta)} \mathbf{1}[Y_{aq} \neq Y_{bq}], \qquad I(q \mid P,\delta) = \frac{L(q \mid P,\delta)}{|C(P,\delta)|}.$$

Influence is the empirical probability that an item separates a close pair. It measures ranking relevance rather than semantic coverage or item quality. Unsupported or degenerate fingerprints are invalid, not artificial zero vectors. Margins are in raw-score units here; a 2.0-accuracy-point margin instead requires $100|S_a - S_b|/Q \le 2.0$.

3.2 Residual influence fingerprints

Population-local pass rate is $D(q \mid P) = |P|^{-1}\sum_{a\in P} Y_{aq}$. Corrected discrimination $A(q \mid P)$ is the correlation between item correctness and the population-local total score with the focal item removed; degenerate correlations are set to zero. Within each population, we fit

$$I(q \mid P) = \beta_0 + \beta_1 D(q \mid P) + \beta_2 D(q \mid P)^2 + \beta_3 A(q \mid P) + \epsilon_q.$$

Residual influence $R(q \mid P) = \epsilon_q$ defines the fingerprint $\mathbf{R}_P = (R(1 \mid P), \ldots, R(Q \mid P))$. All controls and coefficients are estimated locally. This tests what remains after the specified regression; it does not eliminate arbitrary nonlinear effects or interactions.

3.3 Capability populations

The breadth analysis uses target-defined tertiles for diagnostics and *target-excluded capability tertiles* for the primary result. For target benchmark B, model a's capability is its mean standardized score on available non-target benchmarks:

$$Z_a^{(-B)} = |\mathcal{S}_{a,B}|^{-1} \sum_{B'\in\mathcal{S}_{a,B}} z_{aB'}.$$

Here $\mathcal{S}_{a,B}$ contains the non-target benchmarks available for model a; eligibility requires at least two. The target contributes neither responses nor aggregate scores to its population definition. MMLU-Pro uses public BBH (Suzgun et al., 2022) and GPQA performance, with a low/high median split for scaling and three capability tertiles for transfer.

Table 1: Benchmark matrices in the breadth analysis. Eligible models have sufficient non-target coverage for capability stratification.

Benchmark	Models	Items	Eligible
GSM8K	90	997	90
MMLU-Lite	91	565	91
MedQA-Lite	91	998	91
LegalBench-Lite	89	1997	89
Commonsense-Lite	91	500	91
BBQ	42	1000	41
BoolQ	42	3316	41
WikiFact	66	5512	49
ThaiExam	40	560	17

3.4 Fingerprint similarity and estimation

Spearman correlation is the primary fingerprint metric; Kendall's τ_b and top-20, top-50, and top-10% overlaps provide complementary breadth diagnostics. The breadth analysis compares subsamples with full-population references; MMLU-Pro compares fully disjoint panels. Close pairs, controls, regression coefficients, and fingerprints are recomputed within every panel. Large-scale intervals describe repeated panel sampling conditional on the observed model pool (Appendix D).

3.5 Signed influence for held-out transfer

Unsigned influence records whether an item separates models, but discards which model it favors. To select items that preserve ordering, let $C^\star(P, \delta)$ exclude full-score ties and define

$$J(q \mid P, \delta) = \frac{1}{|C^\star(P, \delta)|} \sum_{(a,b) \in C^\star(P,\delta)} \operatorname{sign}(S_a - S_b)(Y_{aq} - Y_{bq}).$$

Positive contributions align with the full-benchmark order; negative contributions oppose it. We residualize J against the same D, D^2, A controls and select items in descending signed residual order. Selected items receive ordinary unweighted scoring, not signed weights. Appendix C.1 checks orientation using scores with the focal item removed. Structural fingerprint analyses continue to use unsigned I.

4 Experimental Setup

The nine-benchmark suite tests breadth; the larger MMLU-Pro matrix supports independent estimation, panel-size analysis, and transfer to unseen models.

4.1 Nine-benchmark breadth suite

We analyze nine cached binary response matrices from the public Fantastic Bugs collection (Truong et al., 2025) (Table 1), retaining complete model rows without imputation. Within each target-excluded capability population, 200 random subsamples retain approximately 90% of models without replacement. Each fingerprint is compared with the full-population reference; between-population similarity compares full-population references. These overlapping comparisons measure subsampling stability. The close-score margin is five correctly answered items. Supporting population and residualization analyses appear in Appendix F.

4.2 Large-scale MMLU-Pro analysis

We retain 441 complete official-provider model rows over 12,032 exactly aligned MMLU-Pro items (Wang et al., 2024). Three of 444 parseable runs have incompatible item sets and are excluded, with no imputation. Integrity checks precede all panel sampling (Appendix A).

Table 2: Target-excluded residual-fingerprint similarity. Within: 90% subsample versus full-population reference. Between: references from different populations. WikiFact has no analyzable cross-population pair.

Benchmark	Within ρ	Between ρ
GSM8K	0.865	-0.020
MMLU-Lite	0.913	-0.018
MedQA-Lite	0.889	0.003
LegalBench-Lite	0.894	0.035
Commonsense-Lite	0.849	-0.008
BBQ	0.925	0.051
BoolQ	0.956	-0.042
WikiFact	0.842	–
ThaiExam	0.942	0.023

Disjoint-panel scaling. The target-excluded capability median split yields 221 low- and 220 high-capability models. For each source population and trial, we draw two non-overlapping panels of size $n \in \{25, 50, 75, 100\}$ and a third panel from the other population. The first panel anchors paired same- and cross-population comparisons. Each direction uses 200 trials; validity requires at least 200 close pairs in every panel under a 2.0-accuracy-point margin.

Held-out transfer. Three 147-model capability tertiles are each split into 73 selection and 74 disjoint evaluation models in each of 200 seeds. The top $K = 50$ items by signed residual influence on selection models are scored by unweighted mean correctness on evaluation models. Held-out close-pair accuracy gives 1 for a preserved order, 0 for a reversal, and $1/2$ for a subset-score tie; full-benchmark ties are excluded. All seeds are valid, with minimum evaluation pair counts $885, 487, 368$ for low, middle, and high populations.

4.3 Additional large-scale target matrices

We construct exact, binary, no-imputation matrices from official-provider Open LLM Leaderboard response files for MuSR (Sprague et al., 2023), IFEval (Zhou et al., 2023), MATH Level 5 (Hendrycks et al., 2021), and GPQA (Rein et al., 2023). Each target has at least 400 aligned models. For every target, capability populations use a median split of standardized public aggregate scores from at least two *other* benchmarks: BBH, GPQA, and MMLU-Pro for MuSR, IFEval, and MATH; BBH and MMLU-Pro for GPQA. We repeat the MMLU-Pro disjoint-panel design with 200 trials per direction, a 2.0-accuracy-point close-pair margin, and at least 200 close pairs in every panel. Curveball uses 100 fixed-margin randomizations. These matrices are distinct target evaluations, although their official-provider model pools overlap; we therefore treat them as corroboration across tasks rather than independent draws of models.

4.4 Fixed-margin null and comparison baselines

Population-wise Curveball randomization preserves every model total and item total while disrupting higher-order response associations (Strona et al., 2014). We apply the same panel and residualization procedures to randomized matrices. For transfer, the primary comparisons use selectors learned from mismatched capability populations as baselines: matched and mismatched selectors are evaluated on the same held-out panel.

5 Population-Specific Structure

5.1 Ranking influence is population-specific across benchmarks

Across all nine breadth matrices, median subsample-to-reference residual-fingerprint correlations range from 0.842 to 0.956, while analyzable cross-population correlations range from -0.042 to 0.051 (Table 2). WikiFact has no analyzable between-population pair.

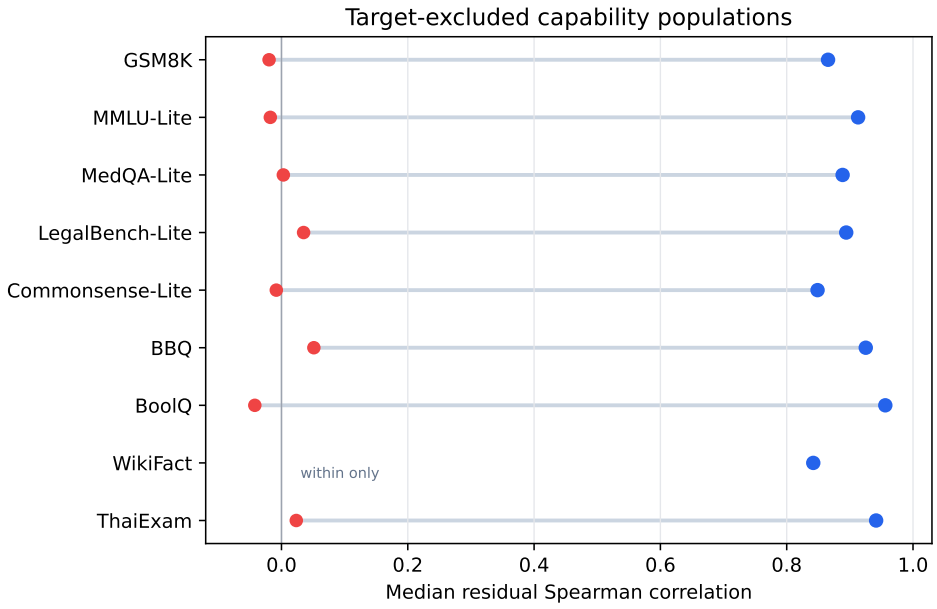

Figure 2: Residual influence across the breadth suite. Within-population stability compares 90% subsamples with full-population references; between-population similarity compares references from different target-excluded capability groups.

Table 3: Disjoint-panel corroboration across additional target benchmarks. Values are median Spearman residual-fingerprint similarities at the listed reliable panel size; gap is the paired same-minus-cross median. Null p95 is the 95th percentile of same-population similarity under fixed-margin Curveball randomization.

Target	Models × Items	n	Same	Cross	Gap	Null p95
MuSR	424 × 756	100	0.252	0.043	0.217	0.051
IFEval	433 × 541	100	0.468	0.186	0.281	0.131
MATH L5	415 × 1324	75	0.249	0.100	0.152	0.079
GPQA	426 × 1192	75	0.220	-0.055	0.271	0.019

This repeated separation appears with target-excluded population labels and after fitting population-local D, D^2, A controls. It establishes breadth of population-conditioned residual structure under overlapping subsampling. Because a subsample shares models with its reference, high agreement does not establish independent recovery. We next test that stronger requirement on MMLU-Pro.

6 Independent Recovery at Scale

6.1 Disjoint model panels recover the same population signature

At $n = 75$, median within-population residual-fingerprint similarity is 0.361, compared with 0.072 across populations. At $n = 100$, the values are 0.435 and 0.081. The median paired same-minus-cross gap widens from 0.288 to 0.349. Each compared pair of panels shares no models, so overlap cannot explain the advantage.

6.2 Additional target tasks corroborate disjoint recovery

Table 3 repeats the disjoint-panel comparison on four target tasks with distinct item sets. GPQA provides a second fully supported scaling result: at $n = 75$, same-population similarity is 0.220, compared with -0.055 across populations, and the Curveball 95th percentile is 0.019. The paired

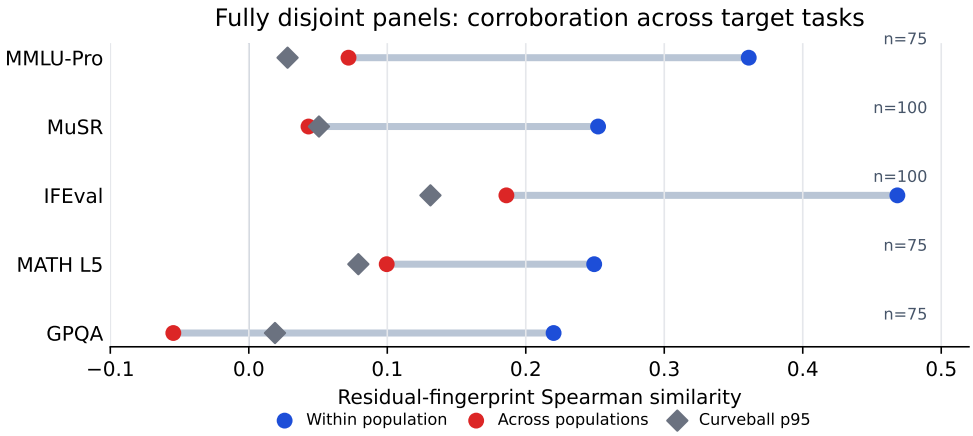

Figure 3: Fully disjoint-panel corroboration across five target tasks. Blue: same-population similarity; red: cross-population similarity; gray diamonds: 95th percentile of same-population similarity under fixed-margin Curveball randomization. Lines connect same- and cross-population values at each target's reliable panel size.

Table 4: Fully disjoint MMLU-Pro recovery. Valid trials require at least 200 close pairs per panel. Gap is the median paired same-minus-cross difference, not the difference of marginal medians. The $n = 50$ row is a low-support diagnostic.

n	Valid	Same ρ	Cross ρ	Gap	Null p95
25	0/400	–	–	–	–
50	10/400	0.357	0.059	0.273	0.015
75	400/400	0.361	0.072	0.288	0.028
100	400/400	0.435	0.081	0.349	0.039

gap is 0.271 with a 95% interval of $[0.260, 0.281]$, and it rises to 0.334 at $n = 100$. MuSR, instruction following, and hard mathematics show the same ordering at their reliable sizes. IFEval has insufficient close-pair support below $n = 100$, and MATH Level 5 has low support at $n = 50$; those support boundaries motivate reporting the reliable panel size for each target rather than forcing a common size.

Small positive cross-population correlations suggest some shared ranking structure. Nevertheless, independent estimates from the same capability regime agree substantially more strongly.

6.3 Population-specific structure becomes recoverable with enough models

At $n = 25$, no trial reaches the minimum close-pair support threshold; at $n = 50$, only 10 of 400 are valid. These settings identify a low-support regime rather than reliable trend evidence. Both $n = 75$ and $n = 100$ are fully valid; the paired-gap bootstrap 95% interval at $n = 75$ is $[0.282, 0.297]$.

Across the two adequately supported sizes, within-population agreement increases more than cross-population agreement. This supports improved recovery of a population-specific component over the tested range. It does not establish behavior at larger, unobserved panel sizes or a universal support threshold across benchmarks.

6.4 Fixed response margins do not explain the population signature

At $n = 75$, observed same-population similarity is 0.361, while the Curveball null's 95th percentile is 0.028. At $n = 100$, these values are 0.435 and 0.039. Preserving model totals and item pass counts does not reproduce the observed agreement: associations in which models succeed on which items carry additional structure.

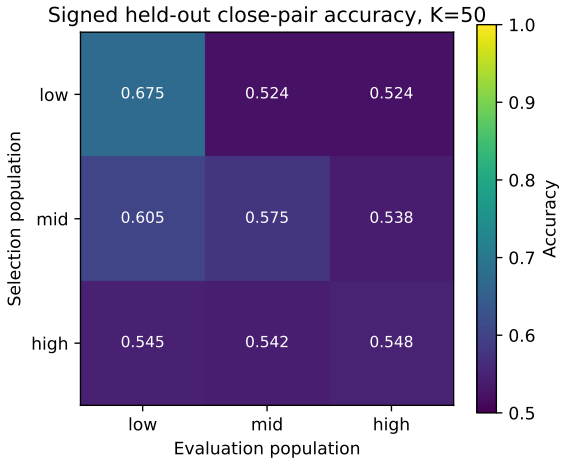

Figure 4: Population-specific signed transfer at $K = 50$. Rows identify selection populations; columns identify disjoint evaluation populations. Each diagonal entry is the best selector within its evaluation column.

7 Population-Specific Transfer

7.1 Population-conditioned influence predicts what matters for unseen models

The structural results motivate a predictive test: does the population in which an item set is learned identify where it will preserve rankings on unseen models? The $K = 50$ transfer matrix is diagonal. Within each held-out evaluation population, the matching selector outperforms both mismatched selectors (Figure 4).

All six matched-minus-mismatched paired effects are positive, with 95% intervals excluding zero (Appendix C). They range from 0.010 for high versus middle selection on high models to 0.130 for low versus high selection on low models. Population labels therefore predict relative ranking utility on models excluded from item selection.

Leave-one-item-out orientation preserves the diagonal pattern and all six positive intervals (Appendix C.1), addressing mechanical self-inclusion of the focal item in the direction statistic. Our downstream claim concerns population-specific transfer rather than universal benchmark-compression optimality.

8 Discussion

Benchmark item importance is relational. An item's ranking relevance can differ across capability regimes without being flawed or invalid. Population-local pass rate and discrimination remain useful summaries, but the specified controls do not exhaust the ranking structure. An influence estimate should therefore identify its reference model population as part of the measurement.

Population conditioning is predictive. The held-out transfer result gives the structural finding practical content: an influence pattern learned from one set of models predicts relative item-set utility for different models from the same regime. This motivates population-aware evaluation and makes claims such as "these questions drive the ranking" meaningful only relative to an explicit population.

Reliable estimation requires model evidence. The MMLU-Pro, IFEval, and MATH results show why panel sizes and qualifying close-pair counts matter. IFEval is reliable only at $n = 100$, while MATH Level 5 has sparse valid $n = 50$ trials; estimates should be interpreted at the sizes with

sufficient close-pair support. Small positive cross-population similarity on some targets also leaves room for shared structure. Separating shared and population-conditioned components is a useful direction for future work.

Ranking relevance and compact benchmark design. Individual influence does not determine an optimal subset: selected items may be redundant or omit useful coverage. Population-specific transfer informs that design problem, but optimizing compact benchmarks may require explicit treatment of interactions among items.

9 LIMITATIONS

Our large-scale evidence spans MMLU-Pro, MuSR, IFEval, MATH Level 5, and GPQA, but these target matrices draw substantially overlapping official-provider model pools from the same evaluation ecosystem. They establish consistency across item domains rather than independence under a new sample of models. Capability is operationalized by target-excluded benchmark performance; grouping models by architecture, training procedure, specialization, or deployment setting would define a different reference population and estimand.

Residualization removes only the variation captured by D, D^2, A, so richer nonlinearities and interactions may remain. The close-pair margin and support threshold specify which comparisons contribute, and the intervals describe stability within the fixed observed model pool rather than uncertainty for future models.

Transfer establishes that matched selectors outperform mismatched selectors, rather than optimal benchmark compression. As a binary response-pattern statistic, influence locates population differences without explaining their causes and does not cover graded scoring, open-ended generation, preference judgments, or interactive evaluation.

10 CONCLUSION

Population-conditioned ranking influence measures relational item importance beyond specified local controls. Across nine matrices, residual fingerprints are population-specific; across five large target matrices, disjoint panels recover the structure beyond fixed response margins; and on MMLU-Pro, signed fingerprints predict item-set utility for held-out models. Thus, **which benchmark items matter for ranking depends on which models are being compared**. Ranking-sensitive analyses should report their reference population and estimator support.

AI USE STATEMENT

Generative AI tools provided limited assistance with manuscript drafting, editing, and implementation support. The authors reviewed and validated all analyses, numerical outputs, figures, and claims, and take full responsibility for the final content. Benchmark responses come from cached model-by-item matrices.

REPRODUCIBILITY STATEMENT

The experiments use cached responses and require no new model inference. Analysis scripts, configurations, cached matrices, result tables, and figures are organized under `scripts/`, `configs/`, `data/cache/`, and `results/`; saved run arguments are under `results/cache/`. Sections 3 and 4 specify the estimators and sampling design. Appendices A, C, and D give alignment checks, paired transfer results, and inferential scope.

REFERENCES

Sang T. Truong, Yuheng Tu, Michael Hardy, Anka Reuel-Lamparth, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin W. Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in AI benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? *arXiv preprint arXiv:2607.15190*, 2026.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations*, 2026.

Zhuo Wang, Wen Wu, Guoqing Wang, Guangze Ye, and Zhenxiao Cheng. MetaEval: Measuring the discrimination of benchmarks for efficient LLM evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33773–33781, 2026.

Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114, 2014.

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*, volume 37, 2024.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS*, 2021.

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. *arXiv preprint arXiv:2311.12022*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

A MMLU-PRO DATA INTEGRITY

The large-scale analyses use an exactly aligned MMLU-Pro response matrix constructed from official-provider runs. The acquisition pipeline identifies 444 parseable runs. Requiring exact item-ID equality retains 441 complete model rows over the same 12,032 items; three runs with incompatible item sets are excluded. No missing responses are imputed.

Before population construction or panel sampling, we verify model-ID uniqueness, exact item ordering, binary response values, matrix dimensions, and capability-source coverage. The resulting $441 \times 12{,}032$ matrix and its integrity checks are fixed for all subsequent large-scale experiments.

B MMLU-PRO PANEL SUPPORT AND NULL DIAGNOSTICS

The disjoint-panel analysis uses a target-excluded capability median split containing 221 low-capability and 220 high-capability models. Every sampled panel independently recomputes close pairs, population-local pass rate, corrected discrimination, residualization coefficients, and residual influence fingerprints.

Under the fixed requirement of at least 200 qualifying close pairs, panel size determines whether fingerprints can be estimated reliably. At $n = 25$, none of the 400 real panel comparisons are valid. At $n = 50$, only 10 of 400 are valid, so this setting is treated only as an estimator-floor diagnostic. Both $n = 75$ and $n = 100$ yield 400 of 400 valid comparisons.

Population-wise Curveball randomization preserves each binary response matrix's row and column totals while disrupting higher-order response associations. The null same-population 95th percentiles are 0.028 at $n = 75$ and 0.039 at $n = 100$, compared with observed same-population medians of 0.361 and 0.435, respectively. At $n = 75$, the observed median paired within-minus-cross gap is 0.288, with bootstrap 95% interval $[0.282, 0.297]$.

C SIGNED TRANSFER DETAILS

Each of the 200 transfer seeds partitions every capability tertile into 73 selection models and 74 disjoint evaluation models. The minimum selection/evaluation counts of non-tied close pairs are $916/885$ for low capability, $472/487$ for middle capability, and $352/368$ for high capability. Thus, all transfer replicates have substantial held-out pair support.

Table 5 reports the six matched-minus-mismatched contrasts for the primary $K = 50$ transfer experiment.

Table 5: Paired population-specific transfer effects at $K = 50$. Each contrast compares the selector learned from the matching capability population with one selector learned from a mismatched population, evaluated on the same held-out model panel.

Evaluation	Mismatched selector	Mean effect	95% interval
Low	Middle	0.0706	[0.0596, 0.0818]
Low	High	0.1302	[0.1170, 0.1430]
Middle	Low	0.0512	[0.0468, 0.0556]
Middle	High	0.0334	[0.0279, 0.0386]
High	Low	0.0245	[0.0188, 0.0303]
High	Middle	0.0104	[0.0040, 0.0169]

All six effects are positive and all six confidence intervals exclude zero. The same matched-population ordering remains visible at $K = 250$:

	eval low	eval middle	eval high
select low	0.678	0.529	0.528
select middle	0.642	0.597	0.558
select high	0.574	0.575	0.571

The $K = 50$ experiment remains the primary downstream analysis.

C.1 LEAVE-ONE-ITEM-OUT SIGNED ORIENTATION

The primary signed statistic orients item disagreements using the full-benchmark ordering of each close model pair. Because that ordering contains the focal item, we test whether the transfer result could arise from mechanical self-inclusion.

We therefore repeat the $K = 50$ transfer experiment using leave-one-item-out orientation,

$$\operatorname{sign}\left(S_a^{(-q)} - S_b^{(-q)}\right),$$

where $S^{(-q)}$ removes item q from the score used to determine ranking direction.

The population-specific diagonal is preserved:

	eval low	eval middle	eval high
select low	0.675	0.524	0.524
select middle	0.601	0.576	0.540
select high	0.547	0.541	0.548

All six matched-minus-mismatched contrasts remain positive, with 95% confidence intervals excluding zero:

$$0.0745\,[0.0632, 0.0853], \quad 0.1288\,[0.1159, 0.1421], \quad 0.0525\,[0.0479, 0.0570],$$

$$0.0354\,[0.0300, 0.0410], \quad 0.0238\,[0.0184, 0.0292], \quad 0.0075\,[0.0011, 0.0135].$$

This check preserves the transfer result after excluding the focal item's contribution from the score used to determine direction.

D STATISTICAL SCOPE

All large-scale confidence intervals and randomization summaries concern repeated panel sampling conditional on the fixed observed 441-model pool. Scaling gaps are paired same-minus-cross correlations within each trial; the reported median of trial-wise differences need not equal the difference of marginal medians. Transfer effects are paired differences in close-pair accuracy on the same evaluation panel, summarized by their mean across seeds; percentile bootstrap intervals resample these seed-level differences. Individual model pairs are not treated as independent replicates. These intervals quantify panel-sampling stability for the observed models, not uncertainty under sampling from an unspecified superpopulation.

E ADDITIONAL RELIABILITY DIAGNOSTICS

The following figures provide complementary diagnostics for the population-conditioned influence estimator. Full Spearman, Kendall, top-20, top-50, and top-10% overlap results are included in the accompanying result files.

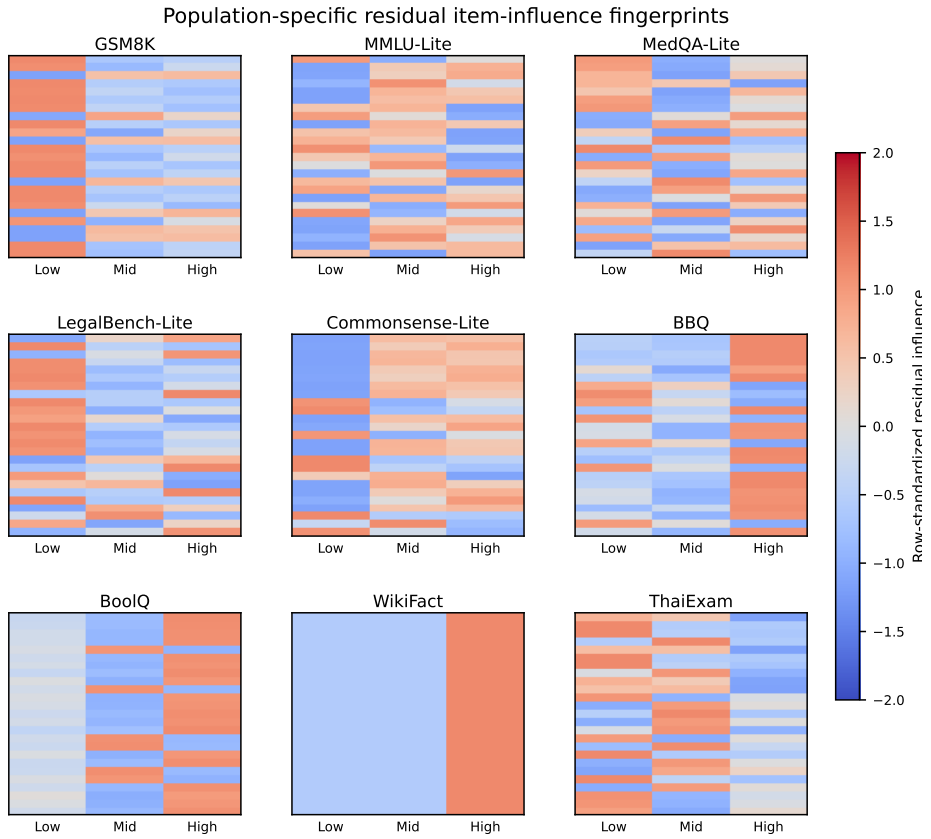

Figure 5: Residual item-by-population influence patterns across the expanded benchmark suite. Rows show high-variance residual-influence items and columns show target-excluded capability populations.

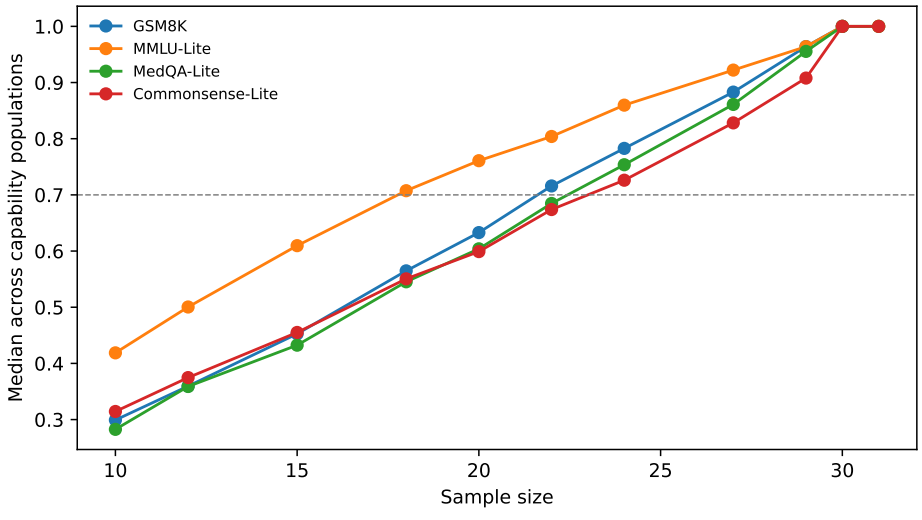

Figure 6: Similarity to the finite full-population residual fingerprint increases as overlapping subsamples contain more unique models.

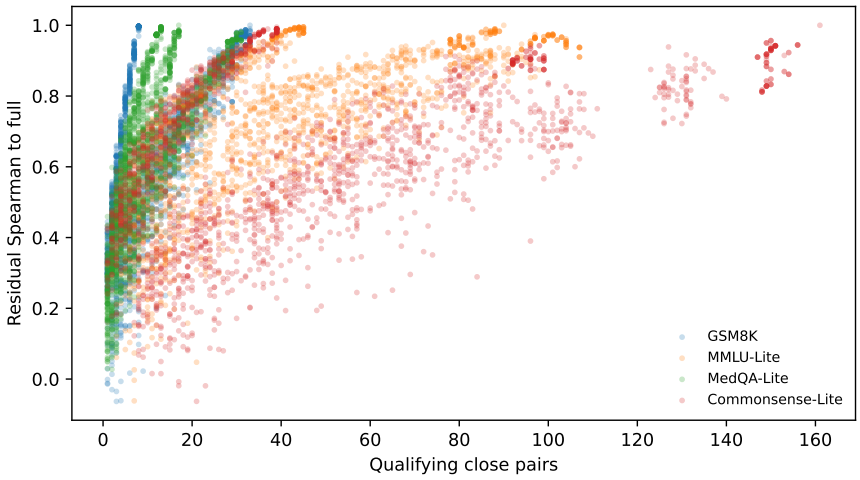

Figure 7: Fingerprint similarity to the finite full-population reference increases with the number of qualifying close model pairs available to the estimator.

F NINE-BENCHMARK SUPPORTING ANALYSES

F.1 COMPLETE TARGET-EXCLUDED BREADTH RESULT

Table 6 reports the complete target-excluded breadth result. Across all nine benchmark matrices, 90% model subsamples closely resemble their full-population residual fingerprints, while correlations between references from different capability populations remain small.

Table 6: Complete target-excluded breadth result across all nine benchmark matrices. Within values compare 90% model subsamples with their full-population reference; between values compare references from different populations.

Benchmark	Within ρ	Between ρ	Difference
GSM8K	0.865	-0.020	0.885
MMLU-Lite	0.913	-0.018	0.931
MedQA-Lite	0.889	0.003	0.886
LegalBench-Lite	0.894	0.035	0.859
Commonsense-Lite	0.849	-0.008	0.857
BBQ	0.925	0.051	0.874
BoolQ	0.956	-0.042	0.998
WikiFact	0.842	–	–
ThaiExam	0.942	0.023	0.918

The repeated separation between high within-population similarity and small cross-population similarity provides breadth for the central phenomenon. The fully disjoint MMLU-Pro experiments in the main paper separately establish that the effect is recoverable without shared models when sufficient model support is available.

F.2 ALTERNATIVE POPULATION DEFINITIONS

Target-excluded capability populations are related to, but not identical to, benchmark-local score tertiles. Table 7 summarizes their agreement before influence is computed.

Table 7: Alignment between benchmark-local score tertiles and target-excluded capability tertiles. Counts include models with both assignments available.

Benchmark	Models	Same-tertile fraction	Rank ρ
GSM8K	90	0.70	0.783
MMLU-Lite	91	0.82	0.912
MedQA-Lite	91	0.74	0.890
LegalBench-Lite	89	0.57	0.666
Commonsense-Lite	91	0.78	0.909
BBQ	41	0.66	0.500
BoolQ	41	0.63	0.640
WikiFact	49	0.63	0.692
ThaiExam	17	0.59	0.922

The assignments are correlated, as expected for two measures of model capability, but are not identical. Crucially, the target benchmark contributes no information to the target-excluded population definition used in the primary breadth analysis.

F.3 RESIDUALIZATION ROBUSTNESS

We finally examine whether reproducible population structure remains after removing population-local pass rate, pass rate squared, and corrected discrimination. Table 8 compares raw and residual within-population correlations from the dedicated robustness analysis.

Table 8: Raw-versus-residual within-population fingerprint stability. Entries are median Spearman correlations from the dedicated residualization robustness analysis.

Benchmark	Raw ρ	Residual ρ	Change
GSM8K	0.789	0.529	0.260
MMLU-Lite	0.919	0.636	0.283
MedQA-Lite	0.800	0.509	0.291
LegalBench-Lite	0.779	0.508	0.271
Commonsense-Lite	0.928	0.533	0.396
BBQ	0.721	0.579	0.142
BoolQ	0.747	0.548	0.199
WikiFact	0.798	0.619	0.179
ThaiExam	0.760	0.540	0.220

Residualization removes a substantial amount of conventional item-level structure, yet reproducible within-population signal remains across all nine benchmarks. The paper's claim therefore concerns the population-specific ranking structure that survives these controls, rather than the weaker observation that raw response patterns differ across capability levels.