LLM BENCHMARK COMPRESSION VIA MAXIMUM INDEPENDENT SET PROMPT SELECTION ON PROXIMITY GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating LLMs across comprehensive benchmarks is expensive, motivating benchmark compression: replacing a full benchmark with a compact subset that is cheaper to run. Whether such a method is usable comes down to fidelity—when does the subset reproduce the full-benchmark LLM ranking, and when does it diverge? We model each benchmark as a prompt proximity graph—nodes are prompts, and an edge joins two prompts whose embeddings are close enough to be treated as semantically redundant—and apply Maximum Independent Set (MIS) solvers to select a maximally diverse, non-redundant subset in which no semantic region is doubly counted. We evaluated four MIS solvers across six embeddings, three distances, six percentile thresholds, and four benchmarks (GPQA, IFEval, MMLU-Pro, Omni-MATH) covering 66 LLMs, spanning 2,558 configurations. Two axes emerge and move almost independently. First, selection is consistent: different random seeds yield near-identical rankings (Kendall's $W \geq 0.90$ in 84.5% of stochastic configurations; mean $W = 0.949 \pm 0.091$), so the procedure itself is reliable. Second, agreement with the full-benchmark ranking is governed by how aggressively the graph is pruned, giving a controllable efficiency–fidelity trade-off: mild thresholds (p_{10}–p_{20}) preserve the full ranking ($\rho = 0.97$–0.98) while removing 64–82% of prompts, whereas aggressive pruning (p_{80}–p_{95}) trades fidelity for compression ($\rho = 0.67$–0.82). Because the axes separate, the grid becomes a recipe: the threshold decides whether a subset stays reliable-and-faithful, and the embedding decides how it fails once it diverges. Divergence concentrates at high thresholds and on the more discretized benchmarks (GPQA, IFEval); when it occurs, a category-level analysis shows MIS redistributes prompt mass toward uniform coverage, which we report descriptively. We recommend ReduMIS with `qwen8b` embeddings, standardized-Euclidean distance, and a p_{20} threshold as a strong default operating point.

1 INTRODUCTION

Rigorous evaluation of large language models (LLMs) requires benchmarks large enough to yield statistically reliable scores, yet small enough to be tractable at scale. Widely used benchmarks such as MMLU-Pro (Wang et al., 2024) and GPQA (Rein et al., 2023) contain hundreds to thousands of questions, incurring substantial inference costs for frontier LLMs and large-scale leaderboards (Polo et al., 2024; Liang et al., 2022). This motivates *benchmark compression*: replacing a full benchmark with a compact subset that is far cheaper to run. The question that decides whether a compression method is usable is one of *fidelity*: when does the subset reproduce the LLM ranking of the full benchmark, and when does it diverge? This paper is organized around characterizing that question so that the choice can be made deliberately.

Our objective is therefore *efficiency*: select a compact, representative subset and quantify how faithfully it preserves the relative ordering of LLMs as a function of how aggressively it compresses. A subset can compress a benchmark with little change in the rankings or with substantial change, and neither is universally preferable. The right choice depends on whether the practitioner values maximal compression or maximal fidelity. A second, *interpretive* question arises whenever a compressed ranking does diverge from the full one: is the divergence merely information lost to pruning, or

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.

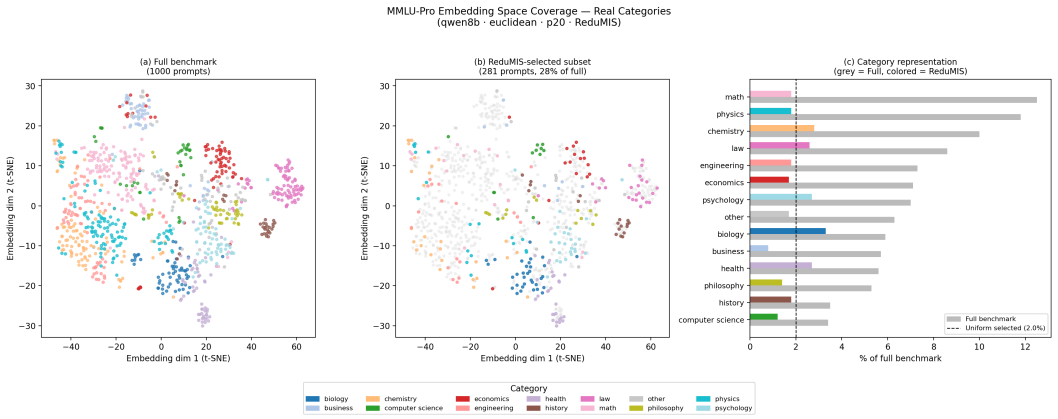


Figure 1: Embedding-space coverage on **real MMLU-Pro categories** (`qwen8b`, standardized Euclidean, p_{20}, ReduMIS). (a) Full benchmark (1000 prompts), coloured by ground-truth category; some categories (e.g. math, physics) are overrepresented. (b) The ReduMIS-selected subset (281 prompts) spreads across the space rather than concentrating in the dense regions. (c) Category proportions: the full benchmark (grey) departs from uniform (2%, dashed), while MIS selection (colored) moves overrepresented categories down and underrepresented ones up.

does it reflect *coverage structure* in the original benchmark? If a benchmark's prompts are unevenly distributed in embedding space—overrepresenting certain semantic regions—then aggregate scores implicitly upweights the capabilities those regions probe; for instance, a reasoning benchmark with disproportionately many similar chemistry questions rewards chemistry ability more than its label suggests (Bohacek et al., 2025). We treat this coverage lens as a way to *describe* what changes under compression, not as a premise that the full benchmark is wrong or that the compressed ranking is more correct. Both the efficiency objective and the coverage lens are the consequence of an identical structural feature of the prompts: *semantic redundancy*. When multiple prompts are close in the embedding space, they *(i)* waste evaluation budget and *(ii)* evaluate the same capability facet across models. Our compression is based on the removal of redundancy; therefore, it also reshapes benchmark prompt distribution. We model each benchmark as a prompt proximity graph $G = (V, E)$, where each prompt is a node $v \in V$ and an edge $(u, v) \in E$ connects two prompts whenever they are close in the embedding space based on different distance thresholds to be considered semantically redundant, i.e. $d(e_u, e_v) \leq \theta$ for a distance threshold θ. A Maximum Independent Set (MIS) of this graph—the largest set of mutually *non-adjacent* nodes—is exactly a maximally non-redundant subset: any two selected prompts satisfy $d(e_i, e_j) > \theta$, so no two selected prompts are semantically close and no semantic region is doubly counted. Figure 2 shows the full pipeline; Figure 1 illustrates the coverage redistribution that accompanies selection on real MMLU-Pro benchmark categories.

This pipeline gives rise to two of our central empirical questions: *(i) how does a compressed subset's agreement with the full-benchmark ranking behave as compression increases, and (ii) is the selection procedure itself reliable?* Agreement with the full-benchmark ranking is not fixed: it is high under mild compression and falls smoothly as pruning becomes aggressive, tracing a *controllable* efficiency–fidelity trade-off. We map this trade-off across the configuration grid so that a practitioner can pick an operating point rather than accept a fixed one, and we separately characterize what changes when the rankings diverge. We find that repeated stochastic MIS selection under different random seeds yields *consistent* rankings—seeds agree closely with each other—so the method is reliable independent of any interpretation. **Contributions: (i)** *A graph-theoretic prompt-selection framework* for benchmark compression that selects a maximally diverse, non-redundant subset via MIS, supporting four MIS solvers, six prompt embedding methods, three distance metrics, and six percentile thresholds, with selected subset size inferred from the graph rather than fixed a priori. **(ii)** *An empirical characterization of the compression–fidelity trade-off* across 2,558 configurations and 66 LLMs on four benchmarks. The two governing axes, random seed concordance and full-benchmark agreement, move almost independently: concordance is high ($W \geq 0.90$ in 84.5% of stochastic configurations), while agreement tracks pruning aggressiveness—mild thresholds (p_{10}–p_{20}) preserve rankings ($\rho = 0.97$–0.98), with divergence concentrated at high thresholds and on

GPQA/IFEval benchmarks. This turns the grid into an operating-point recipe: the threshold sets whether a configuration stays reliable-and-faithful, and the embedding sets how it fails. **(iii)** *Baselines and validation.* Size-matched random and stratified sampling match MIS selection on accuracy (MAE 3.4 points at p_{10}/p_{20}) but drift into the seed-unstable regime under aggressive pruning ($W \to 0.65$–0.79 by p_{90}/p_{95}), isolating MIS's advantage as stability and coverage. A behavioral test confirms the redundancy assumption: retained and discarded neighbors probe comparable difficulty, so the threshold, not the exact distance, identifies redundancy.

2 Related Work

LLM benchmark design and efficiency. LLM benchmarks span reasoning (Wang et al., 2024; Rein et al., 2023), instruction following (Zhou et al., 2023), and advanced mathematics (Cobbe et al., 2021; Hendrycks et al., 2021; Gao et al., 2025), with large suites such as HELM (Liang et al., 2022), Chatbot Arena (Zheng et al., 2023), the Open LLM Leaderboard (Beeching et al., 2023), and LiveBench (White et al., 2024) providing standardized evaluation. Because full-benchmark evaluation is costly, prior work has explored subset selection while preserving ranking fidelity (Perlitz et al., 2023; Vivek et al., 2024). Item Response Theory (IRT) methods use item difficulty, discrimination, or model ability to select informative instances (Polo et al., 2024; Zhou et al., 2026; Truong et al., 2025), while other work studies frontier degradation, benchmark inconsistencies, and multidimensional latent structure (Zhang et al., 2026a; Schilling-Wilhelmi et al., 2025; Chen et al., 2025). These approaches typically estimate model ability on a fixed benchmark and require model interaction. Our approach is complementary and *model-agnostic*: it selects representative prompts from semantic proximity graphs without model evaluations during subset construction. Appendix A (Table 2) further situates our method by selection signal and granularity.

Coverage bias in benchmarks. Overrepresentation of particular question types can amplify the capabilities they probe. Bohacek et al. (2025) show that a small number of concepts dominate coverage in popular benchmarks and disproportionately affect aggregate scores. Other work documents sensitivity to prompt phrasing and annotation artifacts (Mizrahi et al., 2024; Gururangan et al., 2018). We formalize overrepresentation geometrically as prompt density in embedding space: MIS selection redistributes mass away from dense regions because selected prompts cannot be semantically close. We use this descriptively to explain changes under compression rather than assuming that redistribution improves rankings.

Graph-based selection. Graph-based methods such as SELECTOR (Cenikj et al., 2022) construct compact, coverage-preserving subsets, while related approaches have been applied to time-series classification, optimization, and algorithm portfolios (Eftimov et al., 2022; Cenikj et al., 2024; Kostovska et al., 2023). Unlike clustering, MIS determines the subset directly from the graph. We systematically evaluate graph-based semantic prompt selection across six embeddings, three distance measures, six data-driven thresholds, and four MIS solvers on four NLP benchmarks and 66 LLMs.

3 Methodology

Figure 2 shows the full pipeline. Throughout, we describe the construction from the perspective of *semantic redundancy*: two prompts are linked when they are close enough to be semantically redundant based on a particular distance metric, and the selected subset consists of mutually *non-redundant* prompts.

(1) Embedding construction. For each prompt $p_i \in \Pi$ a fixed-dimensional vector is created by an embedding model $f : \mathcal{P} \to \mathbb{R}^d$, $e_i = f(p_i)$. Embeddings place semantically similar inputs near one another, enabling distance-based reasoning about redundancy. Proximity is measured by a distance function d on normalized vectors; larger d means *less* close in the embedding space with respect to the distance metric. **(2) Construction of the proximity graph.** Given embeddings $\{e_1, \ldots, e_N\}$, we build a graph $G = (V, E)$ that makes redundancy explicit. Each prompt p_i is a node v_i. An undirected edge (v_i, v_j) is added whenever the distance between the embeddings of two prompts is at most a threshold, $d(e_i, e_j) \leq \theta$. Therefore, edges connect *redundant* pairs. The threshold θ is set to the p-th percentile of the empirical pairwise-distance distribution, giving a normalized, model-independent knob on graph density. A *small* percentile (p_{10}) yields a *small* θ: only the very closest, most redundant pairs are connected, so the graph is sparse and little is pruned. A *large* percentile (p_{95}) yields a *large*

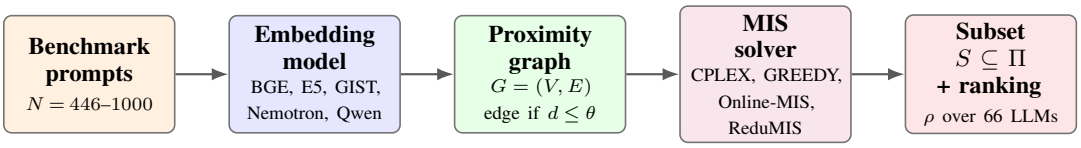


Figure 2: The MIS-based prompt-selection pipeline. Prompts are encoded; a proximity graph is constructed by connecting prompts whose embeddings are within a distance threshold θ (i.e. semantically redundant); an MIS solver selects a maximally diverse, non-redundant subset $S \subseteq \Pi$ of mutually distant prompts for LLM evaluation.

θ: even moderately distant pairs are treated as redundant, the graph is dense, and pruning is aggressive. **(3) Maximum Independent Set.** We cast compression as finding the largest subset $S \subseteq V$ with no edge inside it: for any $p_i, p_j \in S$, $(v_i, v_j) \notin E$, i.e. $d(e_i, e_j) > \theta$. Formally, the solver searches G for an independent set of maximum cardinality, $S^\star = \arg\max_{S \subseteq V} |S|$ subject to $(v_i, v_j) \notin E$ for all $v_i, v_j \in S$ (more details in Appendix B). The selected prompts are therefore pairwise *non-redundant*: no two are within distance θ, so at most one prompt is drawn from any embedding-space region of radius $\theta/2$. Random or stratified sampling does not provide such guarantee. Because the retained prompts jointly cover the space without redundancy, changes in downstream rankings reflect the semantic structure removed by compression, not merely sampling noise. **(4) Evaluation metrics for validation.** To address our first empirical question for each subset S we compute mean LLM accuracy restricted to S, yielding a ranking $\sigma(S)$, and compare it to the full-benchmark ranking $\sigma(\Pi)$ via Spearman's ρ. For descriptive interpretation, we categorized the magnitude of Spearman's rank correlation as negligible ($|\rho| < 0.10$), weak ($0.10 \leq |\rho| < 0.40$), moderate ($0.40 \leq |\rho| < 0.70$), strong ($0.70 \leq |\rho| < 0.90$), and very strong ($|\rho| \geq 0.90$), following Schober et al. (2018). Positive values indicate agreement between rankings, while negative values indicate inverse agreement. In the case of the second empirical question, for stochastic solvers, agreement *across seeds* is measured by Kendall's coefficient of concordance $W = \frac{12\mathcal{S}}{m^2(n^3-n)}$, where m is the number of seeds, n the number of LLMs, and $\mathcal{S} = \sum_i (R_i - \bar{R})^2$ the sum of squared deviations of per-item rank sums; tied per-item ranks are treated as untied. $W \in [0, 1]$, with $W = 1$ perfect agreement. Following Landis & Koch (1977), $W \geq 0.90$ is near-perfect, $W \geq 0.80$ substantial, $W < 0.60$ poor; we use $W \geq 0.90$ as the high-concordance threshold.

4 Experimental Setup

Benchmarks. We evaluate on four benchmarks from the HELM Capabilities leaderboard (Xu et al., 2025) spanning knowledge-intensive QA (MMLU-Pro (Wang et al., 2024) (1000 prompts), GPQA (Rein et al., 2023) (446 prompts)), instruction following (IFEval (Zhou et al., 2023) (541 prompts)), and mathematical reasoning (Omni-MATH (Gao et al., 2025) (1000 prompts)). We restrict to single-turn benchmarks so every instance is an independent text embedding. Table 3 from Appendix C summarizes them across a fixed set of 66 frontier and open LLMs, including GPT (Singh et al., 2025; Agarwal et al., 2025), Claude (Anthropic, 2025), Gemini (Team et al., 2023), Llama (Grattafiori et al., 2024), Qwen2.5/Qwen3 (Qwen Team, 2024; Yang et al., 2025), DeepSeek (Liu et al., 2024), Mistral (Jiang et al., 2023) and OLMo (OLMo Team et al., 2024). Scores were collected and reused from the HELM Capabilities leaderboard; as HELM may reflect differing prompting strategies across families, comparisons are relative to HELM-reported baselines rather than a controlled protocol. MMLU-Pro and GPQA aggregate binary per-item correctness (smooth accuracy, strong rank separability), while IFEval and Omni-MATH produce more discretized distributions; we therefore use rank-based metrics (ρ, W), both robust to discretization. **Sentence embeddings.** We use six models in two regimes: three small encoder-based models (*bge-large-en-v1.5* (Xiao et al., 2024), *e5-large* (Wang et al., 2022), *GIST-small-Embedding-v0* (Solatorio, 2024)) and three large decoder-based models (*Llama-Embed-Nemotron-8B* (Babakhin et al., 2025), *Qwen3-4B*, *Qwen3-8B* (Zhang et al., 2025)). Encoders use a $\sim$512-token window; we apply length-weighted averaging over non-overlapping 510-token chunks. Decoders support 4,096 tokens (sufficient for all prompts; max length 2,278.5 tokens); we additionally evaluate a chunked decoder variant (overlapping windows, stride 256). All models use attention-mask-weighted mean pooling followed by L2 normalization. **Distance measures and thresholds.** We evaluate cosine, Pearson-correlation, and standardized-

Euclidean (*seuclidean*) distances. The threshold θ is a percentile of the empirical pairwise-distance distribution (estimated from 10,000 random pairs); we sweep p_{10} to p_{95}. Percentile thresholds give a model-independent way to control sparsity across embedding spaces with different distance scales (Appendix I). **MIS algorithms.** We use four solvers. **CPLEX** (Anand et al., 2017) solves MIS exactly via an ILP with branch-and-bound/-cut. **GREEDY** iteratively selects the lowest-degree vertex and removes it and its neighbors. **Online-MIS** (Dahlum et al., 2016) is a randomized streaming heuristic (five seeds). **ReduMIS** (Lamm et al., 2017) combines kernelization and stochastic local search (five seeds). On a 1000-node graph, CPLEX needs ∼30–120 s, ReduMIS/Online-MIS up to 600 s, and GREEDY under one second (more details in Appendix L). **Configuration space.** The grid comprises embedding × distance × threshold × solver across four benchmarks; after removing a handful of runs that did not complete, we analyze 2,558 unique configurations; stochastic solvers use five seeds each.

5 RESULTS

We organize the results into several parts. First, we examine selection consistency and the compression–fidelity trade-off, including the behavior of stochastic solvers. We then analyze where divergence is concentrated and compare the proposed methods with the baselines.

Consistency and the compression–fidelity trade-off. Figure 3 places every stochastic configuration (embedding × distance × threshold × stochastic solver across four benchmarks) on the W–ρ plane, gridded by the ρ-magnitude bands of Schober et al. (2018) and the W-concordance bands of Landis & Koch (1977) (the master results file records each configuration's quadrant; see our repository). The two axes move almost independently. Vertically, random seed concordance of the stochastic MIS is high nearly everywhere (mean $W = 0.949 \pm 0.091$; 84.5% near-perfect at $W \geq 0.90$, only 1.9% poor), so a new seed reproduces essentially the same ranking. Horizontally, agreement with the full benchmark spreads across bands, from 53.1% very strong ($\rho \geq 0.90$) down to 1.3% weak or negligible. The two combine most often in the reliable-and-faithful corner ($\rho \geq 0.70$ and $W \geq 0.90$): 77.4% of configurations pair near-perfect seed concordance with strong-or-better baseline agreement. Which quadrant a configuration lands in is set primarily by the threshold—the compression level the practitioner chooses—so reshaping is neither automatically better nor worse than preserving.

Where divergence concentrates. Splitting the plane at $\rho = 0.70$ and $W = 0.90$ gives four quadrants (Table 1), and divergence from the reliable-and-faithful quadrant is unevenly distributed. *By benchmark*, the reliable-and-faithful quadrant holds 89.2% of Omni-MATH and 80.6% of MMLU-Pro configurations but only 64.8% of GPQA and 73.8% of IFEval; the two hardest diverge in complementary ways—IFEval into reliable-but-reshaped (15.4%: stable seeds, $\rho < 0.70$), GPQA into doubly-degraded (14.8%), consistent with its being the smallest, most discretized benchmark. *By threshold*, divergence is a high-percentile effect: $\geq 98.6\%$ of configurations stay reliable-and-faithful through p_{50}, but only 39.7% do at p_{95}, and 100% of $\rho < 0.70$ and 98.5% of $W < 0.90$ cases occur at p_{80}/p_{90}/p_{95}. *By embedding*, large decoders stay on target far more than small encoders (82.5% vs. 67.6%; `nemotron` 88.5% and `qwen8b` 83.8% vs. `gist` 58.3%). These regularities make the trade-off predictable: benchmark, threshold, and embedding together anticipate which quadrant a compression level falls in. Appendix D breaks these outcomes down per benchmark, embedding model, solver, and distance measure (Tables 6–11).

Operating-point recommendations. These patterns turn the plane into a recipe (Table 1): the threshold decides whether a configuration leaves the reliable-and-faithful quadrant, and the embedding decides *how* it fails once it does. Staying at p_{20} or below reaches the target quadrant almost regardless of other choices (100% of p_{10}/p_{20} configurations); when harder compression is required, a large decoder embedding with standardized-Euclidean distance keeps configurations on target (82.5% vs. 67.6% for small encoders, 82.9% vs. 75.5% for cosine; 100% combined with a mild threshold). Our recommended point is `qwen8b`/seuclidean/p_{20}. When a configuration does leave the quadrant, the failure mode is diagnostic and fixable: aggressive thresholds reshape the ranking (lower toward p_{50} if a faithful proxy was intended), weak or small embeddings destabilize seeds (switch to a larger embedding, prefer ReduMIS solver, or average more seeds), and pairing aggressive pruning with weak embeddings on small benchmarks fails on both axes—avoid it (e.g. `gist` @ p_{90}/p_{95} is off-target 93.8% of the time).

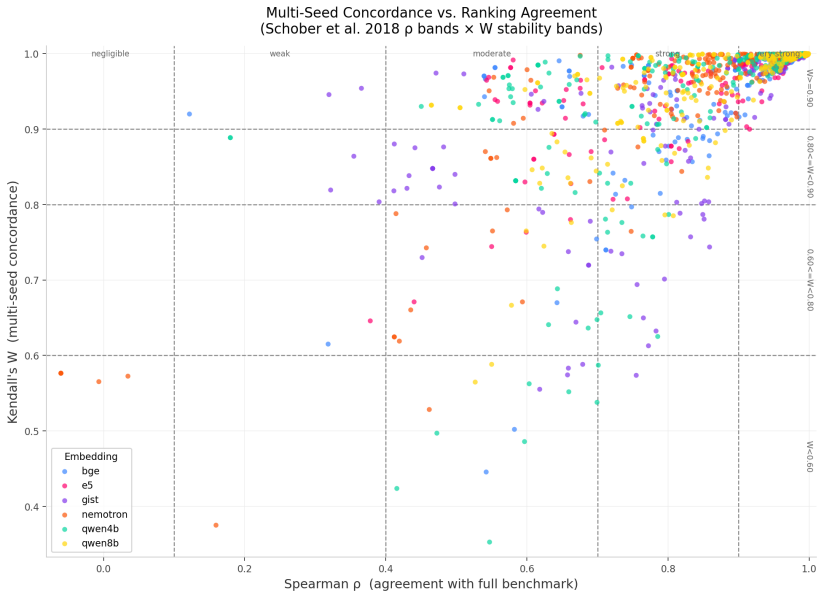


Figure 3: **Multi-seed concordance vs. ranking agreement.** Kendall's W (concordance across seeds) against Spearman's ρ (agreement with the full-benchmark ranking) for every stochastic configuration, colored by embedding model. Dashed grid lines are the interpretation bands defined in Section 3: vertical lines split ρ into the magnitude bands of Schober et al. (2018) (negligible, weak, moderate, strong, very strong, at $0.10/0.40/0.70/0.90$), and horizontal lines split W into the concordance bands of Landis & Koch (1977) ($W < 0.60$ poor; $0.60 \leq W < 0.80$; $0.80 \leq W < 0.90$ substantial; $W \geq 0.90$ near-perfect). Points pile up in the top bands—$W \geq 0.90$ in 84.5% of configurations—so selection is reliable across seeds, while spreading horizontally from the moderate to the very-strong ρ bands: what varies under compression is rank agreement with the full benchmark, not seed stability.

Table 1: Configuration patterns behind each quadrant of Figure 3, and the resulting recommendation. Enrichments ($\times$) are relative to the base rate across all $1{,}262$ stochastic configurations.

Quadrant	**Dominant drivers**	**Recommendation**
Reliable & faithful ($\rho \geq 0.70, W \geq 0.90$; 77.4%)	Mild thresholds (p_{10}/p_{20}: 100%); large decoders; seuclidean	Default target. Use p_{20} or below; for harder compression, large + seuclidean (100% here)
Reliable but reshaped ($\rho < 0.70, W \geq 0.90$; 7.1%)	p_{95}, IFEval, ReduMIS; `qwen4b` at p_{90}–p_{95}	Use deliberately for coverage reshaping + category diagnostic; else lower threshold
Faithful but seed-unstable ($\rho \geq 0.70, W < 0.90$; 7.3%)	Weak/small embeddings: `gist`, `bge`; Online-MIS	Stronger/larger embedding, prefer ReduMIS, or average more seeds
Doubly-degraded ($\rho < 0.70, W < 0.90$; 8.2%)	p_{95}, `gist`, GPQA	Avoid; `gist` @ p_{90}/p_{95} off-target 93.8%. Cap threshold on small benchmarks

Top-N overlap across quadrants. Spearman ρ and Kendall W score the *whole* ranking; a practitioner often cares specifically about the *head* of the leaderboard. To test this, we compute the Jaccard overlap between the top-5 and top-10 models of the full benchmark and those of each subset, and group configurations by the four quadrants explained above. Overlap declines monotonically across the quadrants (Table 5, see Appendix D): the reliable-and-faithful quadrant preserves the head best (top-5 Jaccard 0.559, top-10 0.678; 57% of these configurations share at least half their top-5 and 82% at least half their top-10), and the doubly-degraded quadrant preserves it worst (0.177 / 0.242;

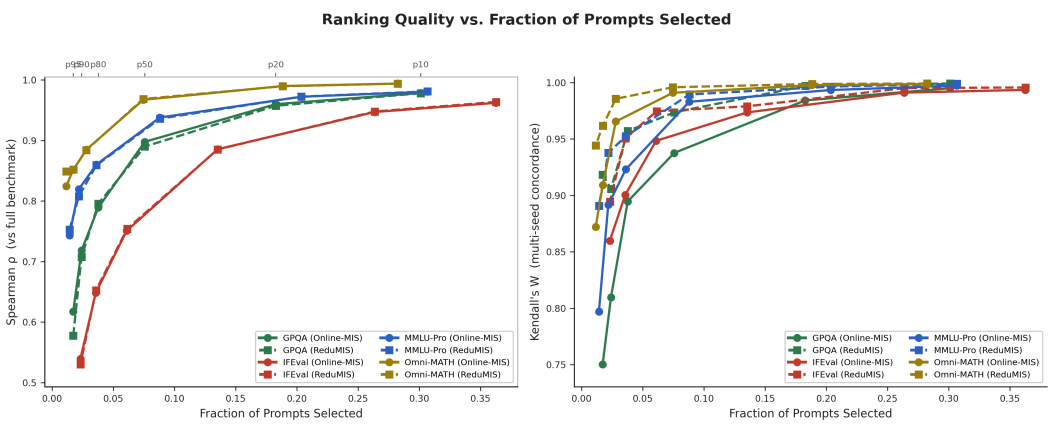


Figure 4: Agreement with the full benchmarks (ρ, left) and seed consistency (W, right) vs. the fraction of prompts selected, per benchmark and stochastic solver. Higher-percentile thresholds sit at the left (small fractions, aggressive pruning). Omni-MATH and MMLU-Pro preserve quality at small fractions; IFEval requires larger subsets.

no configuration reaches half-overlap on top-5, and 18% share no top-5 models at all with the full benchmark). The ordering is not simply a restatement of ρ: the reliable-but-reshaped quadrant ($\rho < 0.70$ but $W \geq 0.90$) keeps more of the head (top-5 0.376) than the faithful-but-seed-unstable quadrant ($\rho \geq 0.70$ but $W < 0.90$; top-5 0.264), showing that seed instability corrupts the top of the leaderboard more than mild whole-ranking reshaping does. As expected, top-10 overlap exceeds top-5 everywhere (the head is a harder target than the shoulder), and overlap falls with the threshold, from 0.693/0.830 at p_{10} to 0.273/0.407 at p_{95}.

Reduction vs. fidelity. Figure 4 shows ρ and W as a function of the fraction of prompts selected, per benchmark and stochastic solver. All benchmarks show fidelity rising with subset size, but with benchmark-specific slopes: Omni-MATH and MMLU-Pro reach high ρ with small fractions, whereas IFEval needs larger subsets for comparable fidelity. The knee of each curve sits near p_{20}, reinforcing it as a robust default. Table 9 from Appendix D quantifies the central trade-off. As the percentile rises from p_{10} to p_{95}, the density of the graph increases, pruning becomes more aggressive, and the mean subset shrinks from ∼231 to ∼11 prompts, increasing the reduction from 64–72% to 98–99%. Fidelity moves in the opposite direction: ρ falls monotonically from 0.979 (p_{10}) to 0.672 (p_{95}) and W from 0.997 to 0.866. The practical consequence is favorable: at p_{20}, the subsets preserve the full ranking well ($\rho = 0.967$, $W = 0.994$) while still removing 74–82% of the prompts. Mild pruning thus captures most of the efficiency benefit with little fidelity cost, while aggressive pruning (p_{80}+) buys marginal extra compression at a steep fidelity price.

Comparison to size-matched baselines. To isolate MIS's contribution beyond size reduction, we compare against random and stratified sampling at *matched* subset sizes—each baseline draws exactly as many prompts as MIS itself selected for that threshold × benchmark cell (averaged over embedding, distance, solver, and chunking), so the subset size is not a free choice for the baselines but one fixed by our own MIS selection. We build three such baselines: random (uniform), stratified (proportional to each benchmark's prompt categories), and uniform-stratified (an equal-per-category share via water-filling), each repeated five times and scored by Spearman ρ and Kendall W across draws. On the same $\rho = 0.70$, $W = 0.90$ plane (Figure 11, Appendix F), all three baselines stay almost entirely in the two faithful ($\rho \geq 0.70$) quadrants but split by seed stability: all three are 100% reliable-and-faithful at p_{10}/p_{20} but collapse by p_{90}/p_{95}, where mean W falls to 0.65–0.79—the mirror image of MIS, which instead holds $W \geq 0.90$ and drifts on ρ under aggressive pruning: baselines keep the ranking but lose seed stability, MIS keeps seed stability but drifts on the ranking. Stratified sampling delays the collapse somewhat (holding reliable-and-faithful through more of the threshold range than random), but even its best case falls far short of MIS's stability, and forcing equal rather than proportional category representation gives back part of that gain rather than extending it. Leaderboard-head overlap follows the same threshold-driven decline: averaged over the five repetitions, top-5 Jaccard falls from 0.75 at p_{10} to 0.32 at p_{95} (top-10: 0.87 → 0.44)(Table 15). Unlike the baselines, MIS

never sees category labels at all: any redistribution it produces is an emergent byproduct of where semantic redundancy concentrates in embedding space, not a targeted allocation, so its coverage shape does not reduce to either baseline's rule. Finally, fixing the subset size automatically as a byproduct of the optimisation is itself an advantage over baselines and clustering methods, which need a cluster count and per-cluster sampling heuristic chosen a priori; Cenikj et al. (2022) analyses this in detail. Full quadrant occupancy, the per-benchmark/per-threshold comparison across all three baselines (Table 16), and category-coverage figures are given in Appendix F.

Absolute-score preservation (MAE). Rank metrics do not capture whether a subset inflates or deflates absolute accuracies (i.e., mean absolute error). For each model we take its mean absolute error over the full benchmark as the reference, compute the mean value on a run's kept subset, and record the signed and absolute error; we average the absolute error over models per run, then over all runs sharing a benchmark and threshold (or a given algorithm)—details in Appendix G. Beyond ranking agreement, subsets also preserve *absolute* accuracy: mean absolute error grows smoothly with compression, from 3.4 accuracy points at p_{10}/p_{20} to 11.3 at p_{90}/p_{95} (Table 19; here and below, p_{10}/p_{20} and p_{90}/p_{95} figures are averaged over the two thresholds in each pair), near-identically across the four solvers. The error is unbiased (mean signed shift $+0.6$ points) and shrinks with subset size, so the aggressive-threshold degradation reflects small-subset variance rather than systematic over- or under-estimation (Fig. 16). Size-matched random subsets track the same curve (mean signed shift -0.3 points) and in fact attain marginally lower MAE at every threshold (2.4 vs. 3.4 at p_{10}/p_{20}, 9.2 vs. 11.3 at p_{90}/p_{95}; Table 19)—consistent with the rank-based comparison, i.i.d. sampling preserves absolute accuracy slightly better while MIS's advantage is seed stability and coverage (Fig. 17). The stratified and uniform-stratified baselines (mean signed shift -0.2 and $+1.1$ points, respectively) both sit between random and MIS (10.2 and 10.7 vs. 9.2 vs. 11.3 at p_{90}/p_{95}; Table 19), so neither proportional nor equal category representation improves absolute-accuracy fidelity over i.i.d. sampling—mirroring its negligible effect on ρ and W (Fig. 18).

Graph topology as a diagnostic. The degree distribution of the full-benchmark graph quantifies how unevenly prompts are distributed: benchmarks with high-degree hubs (many near-duplicate prompts) are where MIS changes coverage most, so a practitioner who cares about coverage can read this off the graph before selecting. Appendix H reports graph-topology statistics across benchmarks and thresholds, giving a heuristic for when a subset can be trusted to preserve the ranking and when the graph is too dense to do so. To identify what governs this topology, we cluster all 648 instances (benchmark $\times$ distance $\times$ threshold) over nine graph-theoretic features and characterize the clusters against the six generative factors. A four-cluster k-means solution is strongly supported (silhouette 0.503) and resolves into a clean 2×2 grid over just two factors: a density regime set by the threshold (sparse, fragmented at p10/p20 vs. near-fully-connected at p50–p95) and a size regime set by the benchmark ($\sim$450–540 vertices for GPQA/IFEval vs. 1000 for MMLU-Pro/Omni-MATH). The rest leave essentially no signature: embedding size, chunking, and embedding model are uniform across clusters, and among distance measures only `seuclidean` perturbs the topology, with `cosine` and `correlation` interchangeable (mean separation 0.002 vs. ~ 0.38 in standardized feature space). Proximity-graph structure is therefore determined almost entirely by edge-retention threshold and benchmark size, largely invariant to embedding and distance-metric choices; full details in Appendix H.

Does embedding proximity imply evaluative redundancy? A quantitative study. Our framework assumes prompts close in embedding space are redundant for evaluation. Rather than rely on human judgment, we test this against model behavior on the three category-labeled benchmarks (MMLU-Pro, GPQA, Omni-MATH), pairing each retained prompt with its nearest discarded neighbor (using *qwen8b*, *seuclidean*, p_{20}). For every prompt we form its *response vector*—the outcomes of all 66 evaluated LLMs—and ask whether embedding-redundant prompts elicit similar behavior along two lines: (i) each prompt's *difficulty*, the fraction of the 66 LLMs that solve it (binarizing graded outcomes at 0.50), and (ii) how much paired response vectors differ (LLMs whose outcome flips, and cosine distance). Together these assess whether the threshold identifies genuine *evaluative* redundancy—prompts on which the field behaves alike—rather than surface proximity; this probes behavioral equivalence, not the stronger claim of identical latent difficulty. Figure 5 asks whether a retained prompt stands in for the difficulty of the neighbor it displaces, placing points along the equal-difficulty diagonal. Pairs are positively correlated on all three benchmarks—weakly on GPQA and MMLU-Pro ($r \approx 0.16$–0.18) and more clearly on Omni-MATH ($r \approx 0.29$), all significant—so a retained prompt probes a comparable difficulty level to the one it replaces, with scatter largest on

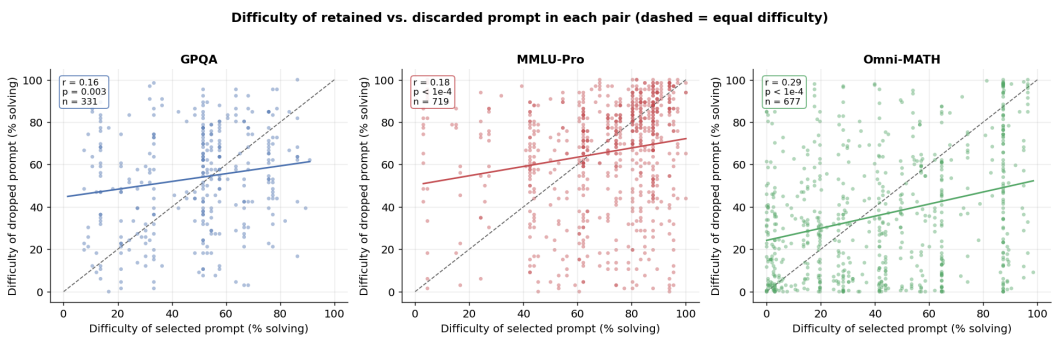


Figure 5: **Difficulty of retained vs. discarded prompt.** For each pair, the discarded prompt's difficulty (% of the 66 LLMs solving it) against its retained neighbour's, per benchmark; the dashed line marks equal difficulty. Points cluster around the diagonal, with the tightest alignment on Omni-MATH. Each panel reports Pearson r, its p-value, and n.

the more discretised benchmarks. The response-vector analysis (Appendix K, Fig. 29) confirms the behavioral reading. If proximity implied redundancy, closer pairs would show higher behavioral similarity; instead the fit is essentially flat within every benchmark ($|r| \leq 0.10$)—so it is the threshold, not the exact distance, that identifies redundancy. Behavioral similarity tracks difficulty rather than geometry: high and tightly peaked for MMLU-Pro (easiest, 68% mean solve rate), lower for GPQA, and broadly spread for Omni-MATH (hardest, 40%), because an easy benchmark yields near-parallel all-correct vectors while a hard one yields sparse, variable ones. Together these support the *behavioral* reading of our assumption: the threshold groups prompts on which the field behaves alike and which sit at comparable difficulty—comparable behavior, not identical latent difficulty.

Category redistribution. Figure 1 grounds the redistribution effect in MMLU-Pro's ground-truth categories rather than a synthetic proxy. Coverage is far from uniform: across 14 categories a uniform split gives 7.1% each, yet math and physics alone account for 12.5% and 11.8%, while history and computer science fall to 3.5% and 3.4% (panel (a)). ReduMIS selection with `qwen8b`, standardized Euclidean distance, and p_{20} retains 281 of 1,000 prompts (panel (b)) and pulls the distribution back toward uniform: math and physics shrink toward 2% while biology and health grow (panel (c)). Appendix E finds the same pattern on GPQA and Omni-MATH. This concretely measures *what* changes when a subset's ranking diverges from the full one: prompt mass is redistributed toward balanced coverage. We report this as a characterization, not a correction — whether flatter coverage is more appropriate than the full-benchmark ranking is an evaluation-goal decision depending on what a practitioner wants the benchmark to measure. **Scope and scalability.** MIS is NP-hard, but solved *once* as offline preprocessing, after which the reduced benchmark is reused at no cost; solvers such as ReduMIS handle large sparse graphs well. Benchmarks with tens of thousands of prompts may need approximate or streaming MIS, which we leave to future work. The framework is text-based here but should transfer to multi-modal or agentic benchmarks by swapping in appropriate embeddings during graph construction; we have not validated this and flag it as a limitation.

6 Conclusion

We introduced an MIS framework for LLM benchmark compression, evaluated across four benchmarks, 66 LLMs, and 2,558 configurations. Selection is highly seed-consistent ($W \geq 0.90$ in 84.5% of configurations), and agreement with the full ranking follows a controllable efficiency–fidelity trade-off: mild thresholds (p_{10}–p_{20}) preserve rankings ($\rho = 0.97$–0.98) while removing 64–82% of prompts. Its edge shows against the baselines: size-matched random and stratified sampling match MIS on accuracy at mild thresholds but destabilize under aggressive pruning ($W \to 0.65$–0.79), whereas MIS stays seed-stable, keeps uniform category coverage, and—unlike sampling—reads its subset size off the graph instead of requiring one fixed a priori. We recommend ReduMIS with `qwen8b`/`seuclidean` at p_{20} as a strong default.

Ethics Statement

This work studies benchmark subset selection over publicly reported evaluation scores and involves no human subjects beyond a small data-driven annotation study of prompt similarity conducted by the authors' team on public benchmark text. We identify three risks. **(1) Adversarial subset selection:** the configurability of the pipeline (embedding, distance, threshold, solver) could be exploited post hoc to favor a particular model; we recommend that benchmark curators fix and publicly disclose the full configuration before evaluation. **(2) Embedding-bias amplification:** MIS redistributes coverage only as captured by the chosen embedding, so cultural, linguistic, or domain biases in the embedding may be amplified rather than reduced, especially for domain-specific or multilingual benchmarks. **(3) Reduced capability coverage:** aggressive pruning may omit rare but important capabilities, which is particularly consequential for safety-critical evaluation; under-represented edge cases should not be discarded for the sake of coverage uniformity. We use only public leaderboard data and release no personal or sensitive information.

Reproducibility Statement

The methodology is fully specified in Section 3, the experimental grid and hyperparameters in Section 5, the embedding-generation and hardware details in Appendix L, and the evaluated models in Appendix C. All benchmark scores are drawn from the public HELM Capabilities leaderboard, and all reported tables and figures are computed from the released per-run statistics. Anonymised source code and the derived per-run data are provided as supplementary material to reproduce every table and figure. Code available at: https://anonymous.4open.science/r/BenchCompress-B6BD/.

Use of AI Assistants

AI assistance was used for grammar and language polishing, as well as auxiliary coding and implementation support that has been checked and verified by the authors; all scientific content, experimental design, and conclusions are the authors' own.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Rimmi Anand, Divya Aggarwal, and Vijay Kumar. A comparative analysis of optimization solvers. *Journal of Statistics and Management Systems*, 20(4):623–635, 2017.

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. Anthropic, 2025. URL https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*, 2025.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open LLM leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2023.

Maty Bohacek, Nino Scherrer, Nicholas Dufour, Thomas Leung, Chris Bregler, and Stephanie C. Y. Chan. Uncovering competency gaps in large language models and their benchmarks, 2025. URL https://arxiv.org/abs/2512.20638.

Gjorgjina Cenikj, Ryan Dieter Lang, Andriess Petrus Engelbrecht, Carola Doerr, Peter Korošec, and Tome Eftimov. SELECTOR: Selecting a Representative Benchmark Suite for Reproducible Statistical Comparison. In *Proc. of Genetic and Evolutionary Computation Conference (GECCO)*, 2022.

Gjorgjina Cenikj, Gašper Petelin, and Tome Eftimov. Transoptas: Transformer-based algorithm selection for single-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '24 Companion, pp. 403–406, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704956. doi: 10.1145/3638530.3654191. URL https://doi.org/10.1145/3638530.3654191.

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 preprint arXiv:2510.00844*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jakob Dahlum, Sebastian Lamm, Peter Sanders, Christian Schulz, Darren Strash, and Renato F. Werneck. Accelerating local search for the maximum independent set problem. In *15th International Symposium on Experimental Algorithms SEA*, volume 9685 of *Lecture Notes in Computer Science*, pp. 118–133. Springer, 2016. URL https://doi.org/10.1007/978-3-319-38851-9_9.

Tome Eftimov, Gašper Petelin, Gjorgjina Cenikj, Ana Kostovska, Gordana Ispirova, Peter Korošec, and Jasmin Bogatinovski. Less is more: Selecting the right benchmarking set of data for time series classification. *Expert Systems with Applications*, 198:116871, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.116871. URL https://www.sciencedirect.com/science/article/pii/S0957417422003189.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Zhengyang Tang, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 100540–100569, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 107–112, 2018.

Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182:105–142, 1999.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Zhongzhan Huang, Guoming Ling, Shanshan Zhong, Hefeng Wu, and Liang Lin. Minilongbench: the low-cost long context understanding benchmark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11442–11460, 2025.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench – a sparse benchmark of reasoning and knowledge in large language models, 2025. URL https://arxiv.org/abs/2407.12844.

Ana Kostovska, Gjorgjina Cenikj, Diederick Vermetten, Anja Jankovic, Ana Nikolikj, Urban Skvorc, Peter Korosec, Carola Doerr, and Tome Eftimov. Ps-aas: Portfolio selection for automated algorithm selection in black-box optimization. In *International Conference on Automated Machine Learning*, pp. 11–1. PMLR, 2023.

Sebastian Lamm, Peter Sanders, Christian Schulz, Darren Strash, and Renato F. Werneck. Finding near-optimal independent sets at scale. *J. Heuristics*, 23(4):207–229, 2017. doi: 10.1007/s10732-017-9337-x. URL https://doi.org/10.1007/s10732-017-9337-x.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Ido Dagan, Leshem Choshen, and Michal Shmueli-Scheuer. Efficient benchmarking of language models. *arXiv preprint arXiv:2308.11696*, 2023.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. Sublime: Subset selection via rank correlation prediction for data-efficient llm evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30572–30593, 2025.

Mara Schilling-Wilhelmi, Nawaf Alampara, and Kevin Maik Jablonka. Lifting the benchmark iceberg with item-response theory. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Aivin V Solatorio. GISTEmbed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*, 2024.

Nishant Subramani, Alfredo Gomez, and Mona T Diab. Simba: Simplifying benchmark analysis using performance matrices alone. In *EMNLP (Findings)*, pp. 13220–13233, 2025.

Maxim K Surkov and Ivan P Yamshchikov. Vygotsky distance: measure for benchmark task similarity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17409–17420, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

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, 2024.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4:2, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Jialiang Xu, Yifan Mai, and Percy Liang. HELM Capabilities: Evaluating LMs capability by capability. Stanford CRFM Blog, March 2025. URL https://crfm.stanford.edu/2025/03/20/helm-capabilities.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Guanhua Zhang, Florian E Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. *Advances in Neural Information Processing Systems*, 38:135638–135678, 2026a.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yueqi Zhang, Jin Hu, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Yiwei Li, Jiayi Shi, Chuyi Tan, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. Learning more from less: Unlocking internal representations for benchmark compression, 2026b. URL https://arxiv.org/abs/2602.00710.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026.

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.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007. doi: 10.4086/toc.2007.v003a006.

Table 2: Where our method sits among benchmark-reduction approaches. Methods differ mainly in the *signal* they consume for selection and the *granularity* at which they select; only embedding-driven methods avoid needing prior model evaluations.

Approach (examples)	**Granularity**	**Selection signal**	**Model evals?**
IRT / anchor items (Polo et al., 2024; Vivek et al., 2024; Kipnis et al., 2025)	sample	model responses	yes
Response-correlation selection (Huang et al., 2025), hidden-state representations (Zhang et al., 2026b), or anchor-LLM + intrinsic features (Saranathan et al., 2025)	sample	model responses / hidden states	yes
Score-based dataset selection (Subramani et al., 2025)	dataset	evaluation scores	yes
Task-similarity selection (Surkov & Yamshchikov, 2024)	dataset	response patterns	yes
Diversity coresets (Abbas et al., 2023)	sample	embeddings	no
MIS on prompt proximity graphs (ours)	sample	embeddings	no[†]

[†]Model scores are used only to *measure* ranking fidelity after selection, never to drive it.

A Positioning Among Benchmark-Reduction Methods

Table 2 summarizes how representative benchmark-reduction methods differ in the signal they consume and the granularity at which they operate, and locates our embedding-only, evaluation-free MIS selection within that landscape. Along the first axis, most prior work is *response-driven*: IRT and anchor-point methods (Polo et al., 2024; Vivek et al., 2024) fit item difficulty and discrimination to a matrix of model responses; methods that instead learn from item-level statistics or model-internal signals — response-correlation matrices (Huang et al., 2025), hidden-state representations (Zhang et al., 2026b), or a small set of anchor LLMs combined with intrinsic sample features (Saranathan et al., 2025) — similarly identify the most informative items; score-based dataset selection (Subramani et al., 2025) compares whole benchmarks using their raw evaluation-score matrices; and task-similarity selection (Surkov & Yamshchikov, 2024) instead compares them by the response patterns they induce across models. All of these require running a pool of models on the full benchmark before a subset can be chosen, and the resulting subset is tied to that pool. Diversity coresets such as SemDeDup (Abbas et al., 2023) and our MIS approach instead use only the prompts' embeddings, so the subset can be constructed once, offline, with no model evaluations, and reused for any future model. SemDeDup has been applied to web-scale image and text training datasets to remove examples based on semantic similarity, using cosine similarity between embeddings combined with a clustering step that requires choosing the number of clusters and a rule for which examples to retain from each group. Unlike clustering, which requires specifying the number of clusters and a rule for selecting examples,MIS instead determines the subset directly from the graph, with no cluster count or retention rule to fix in advance; SELECTOR (Cenikj et al., 2022) presents a direct comparison of MIS-style selection with clustering. Along the second axis, methods operate either at the level of individual samples (selecting prompts within a benchmark) or at the level of whole datasets (selecting benchmarks within a suite); ours is sample-level. What distinguishes MIS from other embedding-only selectors is that the non-redundancy constraint is explicit and the subset size is a by-product of the graph rather than a hyperparameter: coreset and clustering methods must fix a budget or cluster count in advance, whereas the independent-set property guaranties no two retained prompts are within θ and lets the benchmark's own redundancy structure determine how much can be removed. Model scores enter our pipeline only afterward, to measure fidelity (ρ, W, MAE), never to drive selection.

B Maximum Independent Set: Formulation and Algorithm

Let $G = (V, E)$ be the undirected proximity graph, with $V = \{v_1, \ldots, v_N\}$ the prompt nodes and $(v_i, v_j) \in E \iff d(e_i, e_j) \leq \theta$, for a given distance d and a threshold value $\theta \in \mathbb{R}^+$. A set $S \subseteq V$ is *independent* if it contains no edge,

$$\forall\, v_i, v_j \in S,\ i \neq j: \quad (v_i, v_j) \notin E \iff d(e_i, e_j) > \theta. \tag{1}$$

Algorithm 1 Minimum-degree greedy heuristic for MIS

Require: Graph $G = (V, E)$
Ensure: Independent set $S \subseteq V$
1: $S \leftarrow \emptyset$
2: $H \leftarrow G$ ▷ working copy
3: **while** $V(H) \neq \emptyset$ **do**
4: $v \leftarrow \arg\min_{u \in V(H)} \deg_H(u)$ ▷ ties broken arbitrarily/by index
5: $S \leftarrow S \cup \{v\}$
6: remove v and all its neighbours $N_H(v)$ from H
7: **end while**
8: **return** S

Table 3: Benchmark statistics. $|P|$ = prompts; $|M|$ = number of LLMs; Acc. min/max = accuracy range on the full benchmark.

Dataset	|P|	|M|	Acc. min	Acc. max	Top-1 model
GPQA	446	66	0.168	0.803	Gemini 3 Pro (thinking)
IFEval	541	66	0.567	0.951	Grok-3 Mini
MMLU-Pro	1,000	66	0.169	0.903	Gemini 3 Pro (thinking)
Omni-MATH	1,000	66	0.072	0.722	GPT-5 Mini (thinking)

The *maximum independent set* (MIS) is an independent set of maximum cardinality,

$$S^\star \;=\; \arg\max_{S \subseteq V} |S| \quad \text{s.t.} \quad (v_i, v_j) \notin E \;\; \forall\, v_i, v_j \in S. \tag{2}$$

Its cardinality $\alpha(G) = |S^\star|$ is the independence number of G. Equivalently, $S^\star$ is a maximum clique in the complement graph $\bar{G}$, and $V \setminus S^\star$ is a minimum vertex cover of G.

Introducing a binary indicator $x_i \in \{0, 1\}$ for whether $v_i \in S$, Eq. 2 is represented by the binary linear program

$$\max_{x \in \{0,1\}^N} \sum_{i=1}^{N} x_i \tag{3}$$

$$\text{s.t.} \quad x_i + x_j \leq 1 \qquad \forall\, (v_i, v_j) \in E. \tag{4}$$

The edge constraint Eq. 4 forbids selecting both endpoints of any edge, so any feasible x encodes an independent set and the objective counts its size. Solving Eqs. 3–4 exactly yields $S^\star$.

MIS is NP-hard, and hard to approximate within $N^{1-\epsilon}$ for any $\epsilon > 0$ unless P=NP (Håstad, 1999; Zuckerman, 2007). Exact solution via Eqs. 3–4 is therefore tractable only for moderate N or sparse G; for larger or denser graphs we rely on the greedy heuristic below.

When an exact solution is generally impractical to obtain, we may use various approximation techniques, and, among them, standard greedy algorithms take a special place due to their simplicity and wide applicability. The common greedy criterion for solving MIS problem repeatedly selects a minimum-degree vertex and deletes it together with its neighbors (Algorithm 1). Choosing the lowest-degree vertex removes as few candidates as possible at each step. The heuristic at the end returns a *maximal* independent set (no vertex can be added without violating independence) and guaranties $|S| \geq \sum_{v \in V} 1/(\deg(v) + 1)$, although not necessarily the maximum $S^\star$.

The algorithm 1 is deterministic once the tie-breaking rule is fixed; randomized tie-breaking (or randomized solvers generally) induces a distribution over solutions, which motivates the across-seed concordance analysis (Kendall's W).

C Evaluated LLMs

Table 3 summarizes the four benchmarks by prompt count, number of evaluated LLMs, full-benchmark accuracy range, and top-ranked model, showing that GPQA and IFEval are roughly half the size of MMLU-Pro and Omni-MATH and that IFEval has the narrowest accuracy spread.

Table 4: The 66 evaluated LLMs and their providers.

#	Model Key	Provider	#	Model Key	Provider
1	amazon_nova_lite	Amazon	34	gpt_oss_20b	OpenAI
2	amazon_nova_micro	Amazon	35	grok_3_beta	xAI
3	amazon_nova_premier	Amazon	36	grok_3_mini_beta	xAI
4	amazon_nova_pro	Amazon	37	grok_4_0709	xAI
5	claude_3_5_haiku_20241022	Anthropic	38	ibm_granite_3_3_8b_instruct	IBM
6	claude_3_5_sonnet_20241022	Anthropic	39	ibm_granite_4_0_micro	IBM
7	claude_3_7_sonnet_20250219	Anthropic	40	ibm_granite_4_0_small	IBM
8	claude_4_5_haiku_20251001	Anthropic	41	kimi_k2_instruct	Moonshot
9	claude_4_5_sonnet_20250929	Anthropic	42	llama_3_1_instruct_turbo_405b	Meta
10	claude_4_opus_20250514	Anthropic	43	llama_3_1_instruct_turbo_70b	Meta
11	claude_4_opus_..._thinking	Anthropic	44	llama_3_1_instruct_turbo_8b	Meta
12	claude_4_sonnet_20250514	Anthropic	45	llama_4_maverick_17bx128e_fp8	Meta
13	claude_4_sonnet_..._thinking	Anthropic	46	llama_4_scout_17bx16e_instruct	Meta
14	deepseek_r1_0528_with_thinking	DeepSeek	47	marin_8b_instruct	Marin
15	deepseek_v3	DeepSeek	48	mistral_instruct_v0_3_7b	Mistral AI
16	gemini_1_5_flash_002	Google	49	mistral_large_2411	Mistral AI
17	gemini_2_0_flash	Google	50	mistral_small_3_1_2503	Mistral AI
18	gemini_2_0_flash_lite_preview	Google	51	mixtral_instruct_8x7b	Mistral AI
19	gemini_2_5_flash_04_17_preview	Google	52	o3_2025_04_16	OpenAI
20	gemini_2_5_flash_lite	Google	53	o4_mini_2025_04_16	OpenAI
21	gemini_2_5_pro_03_25_preview	Google	54	olmo_2_13b_instruct_nov_2024	AllenAI
22	gemini_3_pro_preview_thinking	Google	55	olmo_2_32b_instruct_mar_2025	AllenAI
23	glm_4_5_air_fp8_with_thinking	Zhipu AI	56	olmo_2_7b_instruct_nov_2024	AllenAI
24	gpt_4_1_2025_04_14	OpenAI	57	olmoe_1b_7b_instruct_jan_2025	AllenAI
25	gpt_4_1_mini_2025_04_14	OpenAI	58	palmyra_fin	Writer
26	gpt_4_1_nano_2025_04_14	OpenAI	59	palmyra_med	Writer
27	gpt_4o_2024_11_20	OpenAI	60	palmyra_x5	Writer
28	gpt_4o_mini_2024_07_18	OpenAI	61	palmyra_x_004	Writer
29	gpt_5_1_2025_11_13	OpenAI	62	qwen2_5_instruct_turbo_72b	Alibaba
30	gpt_5_2025_08_07_with_thinking	OpenAI	63	qwen2_5_instruct_turbo_7b	Alibaba
31	gpt_5_mini_..._with_thinking	OpenAI	64	qwen3_235b_a22b_fp8_thinking	Alibaba
32	gpt_5_nano_..._with_thinking	OpenAI	65	qwen3_235b_a22b_instruct_2507	Alibaba
33	gpt_oss_120b_with_thinking	OpenAI	66	qwen3_next_80b_a3b_thinking	Alibaba

Table 5: Top-N Jaccard overlap between the full-benchmark and subset leaderboards, by quadrant of Figure 3 ($\rho = 0.70$, $W = 0.90$). Overlap is the fraction of shared models among the top 5 (or top 10); "≥ 0.5" is the share of configurations preserving at least half of the head.

Quadrant	top-5 (mean)	top-5 ≥ 0.5	top-10 (mean)	top-10 ≥ 0.5
Reliable & faithful ($\rho \geq .70, W \geq .90$)	0.559	57%	0.678	82%
Reliable but reshaped ($\rho < .70, W \geq .90$)	0.376	26%	0.465	48%
Faithful but seed-unstable ($\rho \geq .70, W < .90$)	0.264	3%	0.399	19%
Doubly-degraded ($\rho < .70, W < .90$)	0.177	0%	0.242	0%

Table 4 lists all 66 LLMs evaluated, drawn from the HELM Capabilities leaderboard.

D How Benchmarks Differ in Subset Characteristics

Table 5 gives the full per-quadrant top-N Jaccard breakdown summarized in Section 5.

Table 6 partitions the stochastic configurations of each benchmark across the four quadrants of Section 5 ($\rho = 0.70$, $W = 0.90$; Figure 3), and Table 7 does the same by embedding model.

Per-dataset analysis. Table 8 reveals clear variation across benchmarks for the stochastic solvers, aggregated across embeddings, distances, and thresholds. Omni-MATH and MMLU-Pro preserve rankings best ($\rho \approx 0.88$–0.92, $W \approx 0.93$–0.98), while GPQA and IFEval are hardest ($\rho \approx 0.79$–0.82), reflecting their smaller size and greater score discretizations. Even on the hardest benchmarks, ReduMIS keeps high seed concordance ($W \geq 0.957$), so subsets remain internally reliable even where they diverge more from the full ranking.

Effect of threshold. Table 9 gives the full per-threshold breakdown underlying the efficiency–fidelity trade-off discussed in Section 5.

Embedding model and distance measure. Table 10 and Figure 6 compare embeddings and distances for the stochastic solvers. `qwen8b` with standardized Euclidean distance is the best configuration

Table 6: Distribution of stochastic subset configurations across the four quadrants of Section 5 ($\rho = 0.70$, $W = 0.90$; Figure 3).

Benchmark	Reliable & faithful ($\rho \geq .70, W \geq .90$)	Reliable but reshaped ($\rho < .70, W \geq .90$)	Faithful but seed-unstable ($\rho \geq .70, W < .90$)	Doubly-degraded ($\rho < .70, W < .90$)
GPQA	64.8%	10.7%	9.7%	14.8%
IFEval	73.8%	15.4%	0.6%	10.2%
MMLU-Pro	80.6%	0.9%	12.0%	6.5%
Omni-MATH	89.2%	1.9%	7.1%	1.9%

Table 7: Per-embedding-model breakdown of stochastic configuration outcomes across the four quadrants of Section 5 ($\rho = 0.70$, $W = 0.90$).

Embedding	Reliable & faithful	Reliable but reshaped	Faithful but seed-unstable	Doubly-degraded
BGE	72.2%	10.4%	13.9%	3.5%
E5	72.2%	12.5%	6.2%	9.0%
GIST	58.3%	5.6%	17.4%	18.8%
Nemotron	88.5%	2.5%	1.4%	7.6%
Qwen-4B	75.0%	10.8%	6.3%	7.8%
Qwen-8B	83.8%	4.6%	6.0%	5.6%

($\rho = 0.921$, $W = 0.984$), closely followed by `qwen4b`+seuclidean ($\rho = 0.901$, $W = 0.982$); for these decoder models the standardized-Euclidean measure is clearly stronger than cosine or correlation. Among small encoders, *e5* and *bge* are the most stable ($W \approx 0.94$–0.96), while *gist* is the weakest overall. Large decoder embeddings give the highest and most consistent scores, but a small encoder (*e5* or *bge*) remains a reasonable choice when GPU memory is limited.

Overall performance by algorithm. Table 11 shows that all four solvers achieve nearly identical Spearman correlation with the full-benchmark ordering ($\rho \approx 0.84$–0.85, aggregated across all benchmarks, embeddings, distances, and thresholds), so global ranking structure is largely invariant to solver choice. Among stochastic solvers, ReduMIS attains the strongest and most stable seed concordance ($W = 0.966$ vs. 0.932 for Online-MIS) with much smaller variance, making it the recommended solver.

E Category Coverage on Additional Benchmarks

Figures 7 and 8 repeat the real-category coverage analysis of Figure 1 for GPQA and Omni-MATH.

Figures 9 and 10 show, for GPQA and Omni-MATH, how each solver redistributes category mass relative to the full benchmark.

F Baselines

Baseline construction. For each tested configuration, defined by the threshold and benchmark dataset, we average across embedding, distance measure, solver, and chunking variant and use that subset size in our baseline calculations. We construct three size-matched baselines: (i) random selection, in which prompts are sampled uniformly from the entire benchmark; (ii) stratified selection, in which prompts are sampled proportionally to each prompt category represented in that benchmark; and (iii) uniform-stratified selection, in which prompts are instead split as equally as possible *across* categories via a water-filling allocation—each category is capped at its own available pool size, and any shortfall from categories that run out is redistributed to categories that still have room—so small and large categories are targeted for comparable representation rather than their natural share of the benchmark. Each baseline is repeated five times to account for sampling stochasticity, and evaluated by Spearman ρ (agreement with the full ranking) and Kendall W across repetitions.

Table 8: Per-dataset, per-algorithm results for the stochastic solvers, aggregated across embeddings, distances, and thresholds. ρ = mean Spearman correlation with the full benchmark; $\rho \geq 0.90$ and $0.70 \leq \rho < 0.90$ = share of seed runs falling in the "very strong" and "strong" Schober bands, respectively; W = mean Kendall concordance; $\tau \geq 0.9$ = fraction of seed pairs with Kendall $\tau \geq 0.9$; $|S|$ = mean subset size.

Dataset	Algorithm	ρ (mean)	ρ_{std}	$\rho \geq 0.90$ (%)	$0.70 \leq \rho < 0.90$ (%)	W	$\tau \geq 0.9$ (%)	$\lvert S \rvert$
GPQA	Online-MIS	0.821	0.155	41.0	37.3	0.8916	40.3	45.4
GPQA	ReduMIS	0.811	0.177	41.5	32.8	0.9572	89.2	45.4
IFEval	Online-MIS	0.789	0.196	39.3	34.8	0.9445	62.6	79.4
IFEval	ReduMIS	0.789	0.205	40.0	34.5	0.9649	86.5	79.5
MMLU-Pro	Online-MIS	0.886	0.107	57.8	35.0	0.9309	65.2	111.6
MMLU-Pro	ReduMIS	0.885	0.107	59.4	33.5	0.9610	86.5	111.7
Omni-MATH	Online-MIS	0.919	0.096	71.1	23.7	0.9557	66.1	100.3
Omni-MATH	ReduMIS	0.923	0.088	71.8	25.1	0.9809	88.4	100.3

Table 9: Effect of the distance threshold, averaged over all solvers and all embedding/distance/benchmark combinations. ρ = mean Spearman correlation with the full benchmark; W = Kendall concordance; $|S|$ = mean subset size; Reduction = relative reduction vs. the full benchmark (min–max across benchmarks). Higher percentiles compress more but preserve rankings less.

Threshold	ρ	ρ_{std}	W	W_{std}	$\lvert S \rvert$	Reduction (%)
p10	0.979	0.015	0.9972	0.0036	230.7	63.9–71.9
p20	0.967	0.024	0.9943	0.0067	154.1	73.8–81.7
p50	0.919	0.059	0.9787	0.0286	67.7	86.6–92.6
p80	0.819	0.117	0.9504	0.0586	28.2	93.9–97.2
p90	0.751	0.146	0.9098	0.1044	17.1	96.5–98.3
p95	0.672	0.195	0.8662	0.1432	11.3	97.7–98.9

Quadrant occupancy. Placing the size-matched baselines on the same $\rho = 0.70$, $W = 0.90$ plane as Figure 3 exposes the difference from MIS structurally (Figure 11): the baselines mostly stay within the two $\rho \geq 0.70$ (faithful) quadrants, but they split between them by *seed stability*. Random selection lands in reliable-and-faithful for 46% of cells and in faithful-but-seed-unstable ($\rho \geq 0.70$, $W < 0.90$) for the other 54%; stratified behaves almost identically (44% / 56%); uniform-stratified is close behind (50% / 44%, with one exception, discussed below). Random and stratified never occupy a $\rho < 0.70$ quadrant—so their failure mode is purely instability, not reshaping—but uniform-stratified does so once, at GPQA p_{95} ($\rho = 0.685$, $W = 0.549$, doubly-degraded): with only 8 prompts split across GPQA's three categories, forcing near-equal representation leaves too little budget per category for a stable ranking. The split is otherwise threshold-driven: at p_{10}/p_{20} all three baselines are 100% reliable-and-faithful, but they diverge in how quickly they collapse as pruning intensifies—stratified stays reliable-and-faithful longest (100% at p_{50}, 67% at p_{80}), while random and uniform-stratified fall off sooner (50%/33% at p_{50}/p_{80}, respectively)—before all three converge to 0% reliable-and-faithful at p_{90}/p_{95}, where mean W falls to 0.65–0.79 (random 0.73, stratified 0.75, uniform-stratified 0.77). This is the mirror image of MIS: where MIS holds $W \geq 0.90$ and drifts on ρ under aggressive pruning (its divergence appears as *reshaping*), the baselines hold ρ but lose W (their divergence appears as *seed instability*). Stratification does not change this picture—it tracks random almost cell-for-cell—confirming that balancing category proportions does nothing for the seed stability that MIS's redundancy removal provides; uniform-stratified confirms this further—forcing *equal* rather than proportional category representation is no more helpful for seed stability, and its head overlap (mean top-5/top-10 Jaccard 0.540/0.637) is statistically indistinguishable from random (0.545/0.649) and stratified (0.542/0.638).

Quadrant distribution, by benchmark and by method. Table 12 and Table 13 give the same four-quadrant breakdown used for MIS in Appendix D (Tables 6–7), computed from the per-configuration ρ and W values underlying Figure 11 (60 configurations total: random n=24; stratified and uniform-stratified n=18 each, undefined for IFEval). Across all 60 configurations, not one lands in the reliable-but-reshaped quadrant ($\rho < 0.70, W \geq 0.90$)—baseline divergence, when it occurs, is exclusively seed instability, never reshaping, confirming the mirror-image relationship with MIS reported above. *By*

Table 10: Mean Spearman ρ and Kendall W by embedding model and distance measure (stochastic solvers). S/L denote small encoder-based and large decoder-based models. **Bold**: best W.

Model	Measure	ρ	W
bge (S)	correlation	0.865	0.9434
bge (S)	cosine	0.861	0.9406
bge (S)	seuclidean	0.844	0.9512
e5 (S)	correlation	0.849	0.9556
e5 (S)	cosine	0.847	0.9576
e5 (S)	seuclidean	0.843	0.9480
gist (S)	correlation	0.791	0.8959
gist (S)	cosine	0.794	0.8952
gist (S)	seuclidean	0.818	0.9065
nemotron (L)	correlation	0.866	0.9634
nemotron (L)	cosine	0.863	0.9624
nemotron (L)	seuclidean	0.863	0.9574
qwen4b (L)	correlation	0.817	0.9314
qwen4b (L)	cosine	0.824	0.9269
qwen4b (L)	seuclidean	0.901	0.9819
qwen8b (L)	correlation	0.857	0.9544
qwen8b (L)	cosine	0.860	0.9555
qwen8b (L)	seuclidean	0.921	**0.9837**

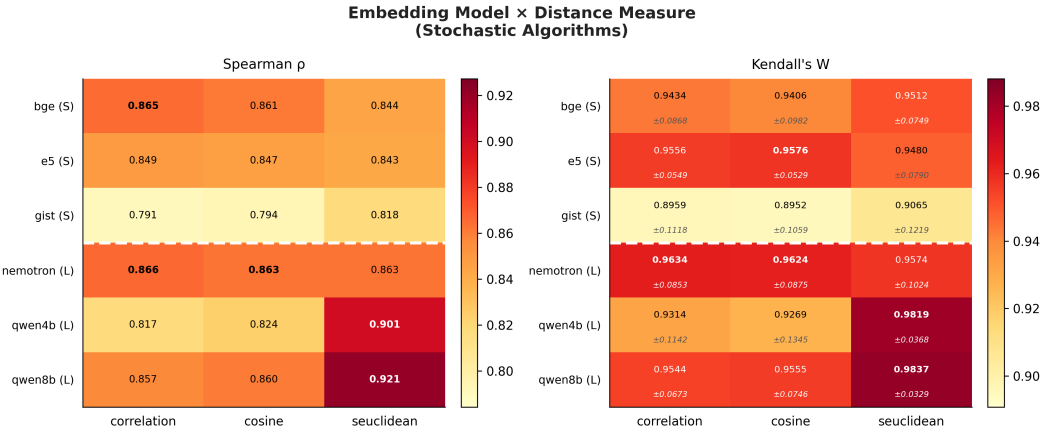


Figure 6: Spearman ρ (left) and Kendall W (right) by embedding model $\times$ distance measure (stochastic solvers). Standardized-Euclidean distance is markedly stronger for the large decoder models (`qwen4b`, `qwen8b`); `qwen8b`+seuclidean is best on both metrics.

benchmark (Table 12), GPQA is hardest for the baselines (38.9% reliable-and-faithful, with 11.1% doubly-degraded—the single random p_{90} boundary case and the single uniform-stratified p_{95} case discussed above), while MMLU-Pro and Omni-MATH are easiest (61.1% each); this mirrors the benchmark-difficulty ordering seen for MIS (Table 6), but at systematically lower occupancy—even the easiest benchmarks keep only 61% of baseline configurations reliable-and-faithful, versus 81–89% for MIS on the same benchmarks. *By method* (Table 13), stratified is the strongest baseline (61.1% reliable-and-faithful, and the only one to entirely avoid the doubly-degraded quadrant), random is weakest (45.8%), and uniform-stratified sits in between (50.0%); the ordering is consistent with stratified's longer hold on the reliable-and-faithful quadrant across thresholds noted above (100% through p_{50}, 67% at p_{80}).

Quadrant drivers. Table 14 gives the baseline analog of Table 1, characterizing what drives each quadrant. As already noted, the reliable-but-reshaped quadrant is empty for every baseline—unlike MIS, which reshapes deliberately under aggressive pruning, a baseline that diverges from

Table 11: Per-algorithm summary, aggregated across benchmarks, embeddings, distances, and thresholds. ρ = mean Spearman correlation with the full-benchmark ranking; $\rho \geq 0.95$ = fraction of configurations with high rank agreement; W = mean Kendall concordance across seeds; W_{std} = its standard deviation; $|S|$ = mean selected subset size. W is undefined (—) for the deterministic solvers CPLEX and GREEDY. **Bold**: best stochastic concordance.

Algorithm	ρ (mean)	$\rho \geq 0.95$ (%)	W	W_{std}	$\lvert S \rvert$
CPLEX	0.853	39.0	—	—	84.8
GREEDY	0.843	37.7	—	—	82.5
Online-MIS	0.854	38.0	0.9317	0.1069	85.2
ReduMIS	0.853	38.0	**0.9662**	**0.0675**	85.3

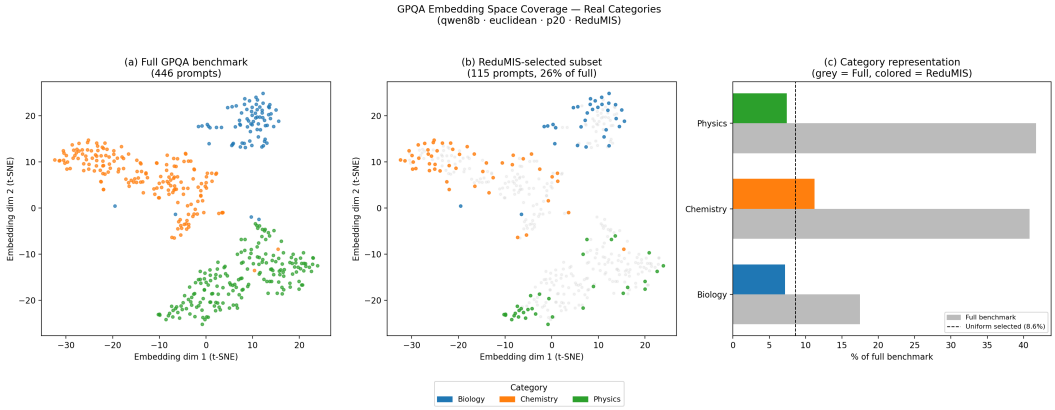


Figure 7: Real-category embedding-space coverage for GPQA (`qwen8b`, standardized Euclidean, ReduMIS).

the full ranking never does so while staying seed-stable; its divergence is instability, full stop. The reliable-and-faithful quadrant is dominated by mild thresholds (p_{10}/p_{20} alone are 65% of it, each enriched $1.94\times$ over the base rate) and is reached slightly more often by stratified sampling than by the other two methods. The faithful-but-seed-unstable quadrant mirrors this at the other end of the threshold range (p_{90}/p_{95} together are 67% of it, each enriched $2.00\times$), with IFEval over-represented ($1.48\times$) and random the most common method there. The tiny doubly-degraded quadrant (3.3%, 2 configurations) is confined entirely to GPQA at its two most aggressive thresholds ($3.33\times$ enrichment) and never occurs under stratified sampling—small, category-imbalanced benchmarks pushed to extreme compression are where a size-matched baseline is most likely to fail outright.

Leaderboard-head overlap (Jaccard). Leaderboard-head overlap declines the same way: averaged over the five repetitions, top-5 Jaccard falls from 0.75 at p_{10} to 0.32 at p_{95} (top-10: $0.87 \rightarrow 0.44$), with all three baselines indistinguishable on this measure (0.534 vs. 0.538 vs. 0.540 on top-5). Grouped by quadrant (Table 15), the faithful quadrants separate by stability—reliable-and-faithful preserves the head best (top-5 mean 0.701, top-10 0.776; 84% of these configurations share at least half their top-5 and 97% at least half their top-10), whereas seed-unstable degrades it (top-5 mean 0.388, top-10 0.514; only 19% reach half-overlap on top-5)—so even at high ρ, an unstable subset ($W < 0.90$) still scrambles the top. Within a quadrant the baselines edge MIS on head overlap (e.g. 0.701 vs. 0.559 on top-5 in reliable-and-faithful), consistent with their small ρ advantage: the trade-off is as above—baselines win marginally on ranking, MIS on stability. Forcing equal rather than proportional category representation (uniform-stratified) is no more helpful for seed stability, and it departs further from the full benchmark's own category composition than stratified does, at no gain in ranking fidelity or head overlap.

Compression vs. quality, all four methods. Figure 12 plots Spearman ρ (left) and the stratified/uniform-stratified category-deviation metric (right) against threshold for all benchmarks with category labels, extending Figure 11 with the same random/stratified/uniform-stratified/MIS grouping used in Table 16. The left panel shows the four methods follow each other closely at mild

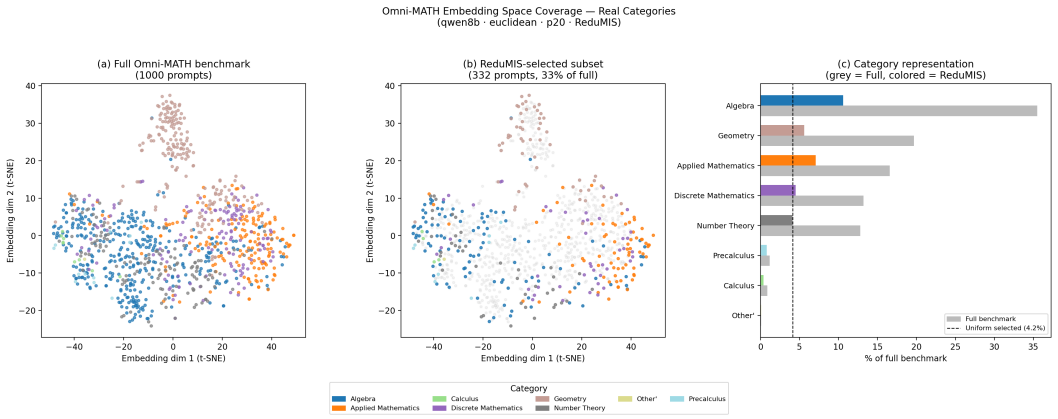


Figure 8: Real-category embedding-space coverage for Omni-MATH (`qwen8b`, standardized Euclidean, ReduMIS).

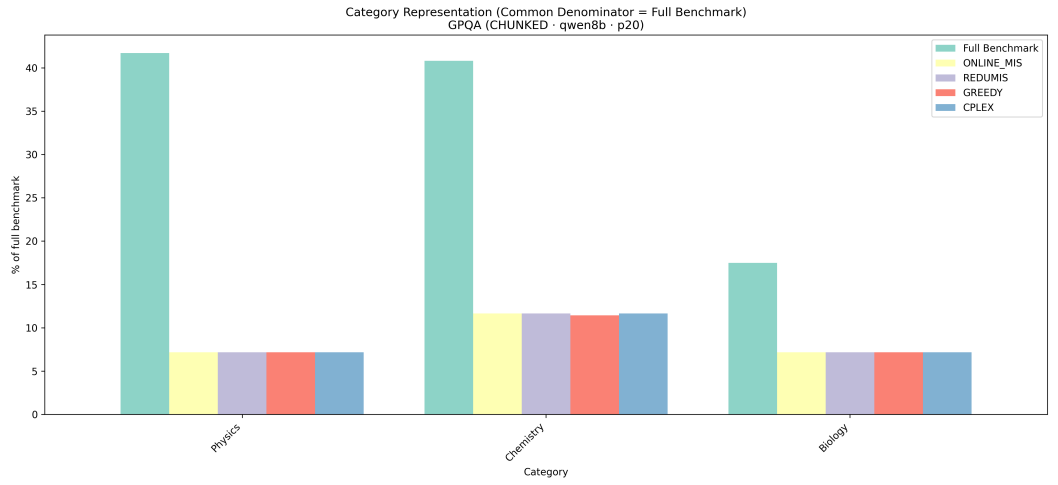


Figure 9: Category proportions on GPQA (`qwen8b`, p_{20}): full benchmark vs. the four MIS solvers.

thresholds ($\rho \geq 0.93$ for all methods at p_{10}/p_{20} on every benchmark) and separate only once pruning becomes aggressive, where MIS is the method that drifts lowest in ρ (its divergence is reshaping, not instability, as established in the main text). The right panel makes explicit what the two stratified variants are optimizing for: stratified's deviation from the full benchmark's own category proportions stays small and roughly flat across thresholds (under 5 points once thresholds narrow the subset, since its explicit goal is to preserve those proportions), whereas uniform-stratified's deviation from the full proportions is larger and does not shrink with threshold, because its allocation target is deliberately *not* the full benchmark's proportions but an equal split across categories. The size of that gap tracks how imbalanced each benchmark's categories are: on MMLU-Pro, whose 14 categories are relatively similar in size, the deviation stays modest (3–5 points); on GPQA and Omni-MATH, whose few categories are far more imbalanced (GPQA is 42%/41%/17%; Omni-MATH's largest category alone is 35%), the deviation is much larger (8–16 points on GPQA, 17–21 points on Omni-MATH)—a large deviation from full there is expected behavior, not degradation.

Category coverage under compression. Figures 13–15 repeat the real-category coverage comparison of Figure 1 at p_{20} for GPQA, MMLU-Pro, and Omni-MATH, this time contrasting all three size-matched baselines against the full benchmark, with every bar expressed as a percentage of the *full* benchmark size N so that a method's bars for a given category are directly comparable to that category's true share (the full benchmark's bars, by construction, sum to 100%; each baseline's bars sum to roughly its subset fraction of N, e.g. $\approx 18\%$ for GPQA at p_{20}). Within that shared

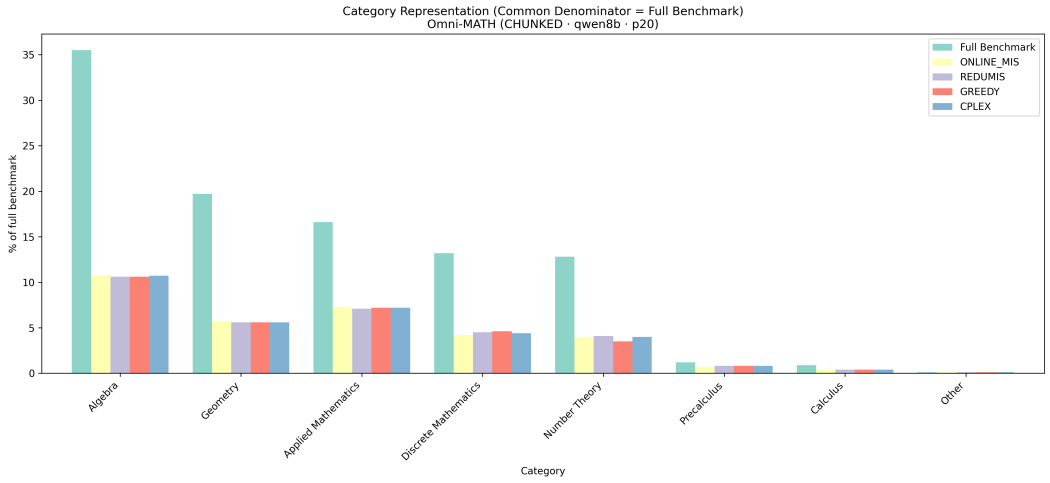


Figure 10: Category proportions on Omni-MATH (`qwen8b`, p_{20}): full benchmark vs. the four MIS solvers.

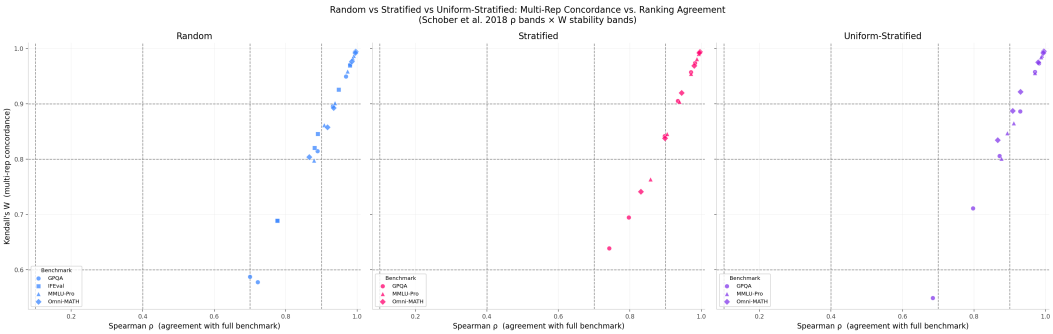


Figure 11: **Size-matched baselines on the concordance–agreement plane.** Multi-rep Kendall's W (concordance across the five repetitions) against Spearman's ρ (agreement with the full benchmark) for the random (left), stratified (middle), and uniform-stratified (right) baselines, one point per (benchmark, threshold) cell, marker-coded by benchmark. Grid lines are the same interpretation bands as Figure 3: the ρ-magnitude bands of Schober et al. (2018) ($0.10/0.40/0.70/0.90$) and the W-concordance bands of Landis & Koch (1977) ($0.60/0.80/0.90$). All three baselines occupy almost exclusively the faithful ($\rho \geq 0.70$) side of the plane and slide *downward* in W as the threshold grows—from near-perfect concordance at p_{10}/p_{20} (top-right) into the seed-unstable band ($W < 0.90$) by p_{80}–p_{95}—so their divergence from the full ranking mostly takes the form of *repetition instability*, not reshaping. This is the mirror image of MIS (Figure 3), which retains $W \geq 0.90$ and instead drifts leftward in ρ. Stratified and uniform-stratified track random almost cell-for-cell, with one exception: GPQA at p_{95} under uniform-stratified sampling is the single point that falls below $\rho = 0.70$ (bottom left of the right panel), where an 8-prompt subset split near-equally across GPQA's three categories is too small for a stable ranking.

subset budget, stratified (purple) reproduces the full benchmark's *relative* category shares—GPQA's Physics/Chemistry/Biology split of roughly $42\%/41\%/17\%$ of the full benchmark is mirrored in stratified's subset composition at essentially the same ratio, so its bars scale down uniformly from the full-benchmark bars without changing their relative heights. Uniform-stratified (red) instead flattens every benchmark toward an *equal* per-category share of the subset regardless of the category's true size: on GPQA this means Physics, Chemistry, and Biology each receive close to one-third of the subset even though Biology is barely a sixth of the full benchmark, visibly lifting Biology's bar above stratified's and pulling Physics/Chemistry's bars below it (Figure 13). The effect is most dramatic on Omni-MATH, where Algebra alone is 35% of the full benchmark and Precalculus and Calculus are each under 2%: uniform-stratified pushes all eight categories toward an equal 12.5%-of-subset

Table 12: Distribution of baseline configurations across the four quadrants of Section 5 ($\rho = 0.70$, $W = 0.90$; Figure 11), by benchmark, pooled across the three baseline methods. Stratified and uniform-stratified are undefined for IFEval, so its n reflects random only.

Benchmark	**Reliable & faithful** ($\rho \geq .70, W \geq .90$)	**Reliable but reshaped** ($\rho < .70, W \geq .90$)	**Faithful but seed-unstable** ($\rho \geq .70, W < .90$)	**Doubly-degraded** ($\rho < .70, W < .90$)	n
GPQA	38.9%	0.0%	50.0%	11.1%	18
IFEval	33.3%	0.0%	66.7%	0.0%	6
MMLU-Pro	61.1%	0.0%	38.9%	0.0%	18
Omni-MATH	61.1%	0.0%	38.9%	0.0%	18

Table 13: Per-baseline-method breakdown of configuration outcomes across the four quadrants of Section 5 ($\rho = 0.70$, $W = 0.90$), pooled across benchmarks.

Method	**Reliable & faithful**	**Reliable but reshaped**	**Faithful but seed-unstable**	**Doubly-degraded**	n
Random	45.8%	0.0%	50.0%	4.2%	24
Stratified	61.1%	0.0%	38.9%	0.0%	18
Uniform-stratified	50.0%	0.0%	44.4%	5.6%	18

target, compressing Algebra's bar down and visibly lifting Precalculus and Calculus above both the full-benchmark and stratified bars (Figure 15); the same flattening, though milder, is visible across MMLU-Pro's longer, more balanced tail of 14 categories (Figure 14). This is the category-level counterpart of the ρ/W result above: uniform-stratified buys deliberate, controllable coverage of minor categories at the cost of departing further from the full benchmark's own composition than stratified does, while neither approach improves seed stability over random sampling.

Adding MIS's own selection (blue) to the same panel requires one caveat: the baselines' bars are sized to the subset size *averaged across all six embeddings and three distances* (Appendix F), whereas MIS's bars use the size actually returned by its best-performing configuration, `qwen8b`+standardized-Euclidean/ReduMIS, which is meaningfully larger at p_{20}—116 vs. 82 on GPQA, 284 vs. 204 on MMLU-Pro, and 332 vs. 187 on Omni-MATH (1.4–1.8×)—because that embedding/distance pair yields a sparser prompt proximity graph at the same percentile threshold and so prunes less aggressively there. Part of MIS's higher bars in Figures 13–15 is simply this larger subset; even so, the *shape* of the redistribution is informative. On GPQA, MIS lifts Chemistry and Biology well above their proportional share—on Chemistry even further than uniform-stratified's deliberate equalization—while Physics tracks closer to its natural share. On Omni-MATH, MIS similarly favors Geometry, Applied Mathematics, and Discrete Mathematics relative to Algebra's outsized weight in the full benchmark, and on MMLU-Pro it lifts several mid-sized categories (Chemistry, Law, Psychology, Biology, Health) while compressing the two largest, Math and Physics. Unlike stratified and uniform-stratified, MIS never sees the category labels at all: this redistribution is an emergent byproduct of where semantic redundancy concentrates in embedding space, not a targeted allocation, so its shape does not reduce to either baseline's rule.

G Absolute-Score Preservation (MAE)

Subset accuracy error (MAE). Beyond ranking agreement, we measure how closely a subset reproduces each model's *absolute* accuracy. For a given benchmark (GPQA, IFEval, MMLU-Pro, Omni-MATH), the baseline is each model's mean accuracy over the *full* prompt set—the average of that model's per-prompt results. Every solver run is defined by an algorithm, chunking variant, benchmark, threshold, seed, and distance measure, and yields a subset of retained instances. For each model m we recompute its mean accuracy on that subset, $\mathrm{acc}_m^{\mathrm{subset}}$, and compare it to its full-benchmark accuracy $\mathrm{acc}_m^{\mathrm{full}}$, giving one signed error and one absolute error per model. Averaging the

Table 14: Configuration patterns behind each quadrant for the size-matched baselines (Figure 11), analogous to Table 1. Enrichments ($\times$) are relative to the base rate across all 60 baseline configurations (random n=24; stratified/uniform-stratified n=18 each, undefined for IFEval).

Quadrant	Dominant drivers	Takeaway
Reliable & faithful ($\rho \geq 0.70, W \geq 0.90$; 51.7%)	Mild thresholds (p_{10}/p_{20}: $\times 1.94$ each, 65% of quadrant); MMLU-Pro/Omni-MATH ($\times 1.18$); stratified slightly favoured ($\times 1.18$)	Baselines are usable proxies only at mild compression; stratified is the safest default among the three
Reliable but reshaped ($\rho < 0.70, W \geq 0.90$; 0.0%)	Not observed (0/60)	Baselines never reshape the ranking while staying seed-stable—divergence, when it happens, is always instability, the mirror image of MIS
Faithful but seed-unstable ($\rho \geq 0.70, W < 0.90$; 45.0%)	Aggressive thresholds (p_{90}/p_{95}: $\times 2.00$ each, 67% of quadrant); IFEval ($\times 1.48$); random slightly favored ($\times 1.11$)	Expect instability once compression is aggressive; repeat draws (especially for random) rather than trusting a single sample
Doubly-degraded ($\rho < 0.70, W < 0.90$; 3.3%)	p_{90}/p_{95} only ($\times 3.00$ each); GPQA exclusively ($\times 3.33$, 100% of quadrant); never stratified ($\times 0.00$)	Confined to GPQA's most aggressive thresholds, where the smallest, most category-imbalanced benchmark leaves too few prompts per category; avoid random/uniform-stratified there

Table 15: Top-N Jaccard overlap between the full-benchmark and size-matched baseline subset leaderboards, by quadrant of Figure 11 ($\rho = 0.70$, $W = 0.90$). Overlap is the fraction of shared models among the top 5 (or top 10); "≥ 0.5" is the share of configurations preserving at least half of the head.

Quadrant	top-5 (mean)	top-5 ≥ 0.5	top-10 (mean)	top-10 ≥ 0.5
Reliable & faithful ($\rho \geq .70, W \geq .90$)	0.701	84%	0.776	97%
Reliable but reshaped ($\rho < .70, W \geq .90$)	0.000	0%	0.000	0%
Faithful but seed-unstable ($\rho \geq .70, W < .90$)	0.388	19%	0.514	44%
Doubly-degraded ($\rho < .70, W < .90$)	0.183	0%	0.291	0%

per-model absolute errors over all N_{models} models summarizes a single run:

$$\text{run MAE} \;=\; \frac{1}{N_{\text{models}}} \sum_{m=1}^{N_{\text{models}}} \left| \, \text{acc}_m^{\text{subset}} - \text{acc}_m^{\text{full}} \, \right|. \tag{5}$$

The signed error is retained separately to expose systematic over- or under-estimation, whereas Eq. equation 5 captures the typical magnitude of the discrepancy. To report a stable figure we average the run-level MAE across all N_{runs} runs that share a configuration—either all runs with the same benchmark and threshold, or all runs of a given algorithm:

$$\text{mae_mean} \;=\; \frac{1}{N_{\text{runs}}} \sum_{r=1}^{N_{\text{runs}}} \text{run MAE}_r. \tag{6}$$

Lower values indicate that the subset preserves not only the *ordering* of models (captured by ρ and W) but also their absolute accuracy levels.

Absolute-accuracy error. Figure 16 reports how well subsets reproduce each model's full-benchmark accuracy, complementing the rank-based analysis. The error is governed by the compression level, not the solver: averaged over all configurations, the four solvers differ by less than 0.2 accuracy points (CPLEX 7.31, GREEDY 7.33, Online-MIS 7.29, ReduMIS 7.18), whereas moving from mild to aggressive thresholds more than triples the MAE; Table 17 gives the full per-benchmark, per-threshold breakdown. Panels (a)–(b) show this degradation is broad—by p_{95} roughly half the

Table 16: Full comparison of MIS against the three size-matched baselines (random, stratified, uniform-stratified), by benchmark and threshold. ρ = mean Spearman correlation with the full ranking; W = mean Kendall concordance across repetitions (5 seeds for MIS/random, 5 reps for stratified/uniform-stratified). Stratified and uniform-stratified require category labels and are undefined for IFEval (–).

			MIS		Random		Stratified		Uniform-Stratified	
Dataset	Threshold	$\lvert S\rvert$	ρ	W	ρ	W	ρ	W	ρ	W
GPQA	p_{10}	134	0.978	0.997	0.981	0.971	0.982	0.974	0.981	0.973
	p_{20}	82	0.959	0.991	0.969	0.949	0.972	0.956	0.967	0.955
	p_{50}	34	0.894	0.956	0.932	0.896	0.930	0.883	0.929	0.898
	p_{80}	17	0.792	0.926	0.889	0.815	0.854	0.773	0.831	0.723
	p_{90}	11	0.713	0.858	0.700	0.587	0.799	0.711	0.802	0.696
	p_{95}	8	0.597	0.834	0.722	0.578	0.743	0.607	0.743	0.607
IFEval	p_{10}	196	0.963	0.995	0.980	0.970	–	–	–	–
	p_{20}	142	0.947	0.993	0.948	0.925	–	–	–	–
	p_{50}	73	0.885	0.976	0.890	0.846	–	–	–	–
	p_{80}	33	0.752	0.961	0.881	0.820	–	–	–	–
	p_{90}	19	0.650	0.926	0.777	0.688	–	–	–	–
	p_{95}	12	0.535	0.877	0.776	0.689	–	–	–	–
MMLU-Pro	p_{10}	307	0.981	0.998	0.990	0.987	0.992	0.989	0.990	0.988
	p_{20}	204	0.972	0.995	0.987	0.982	0.988	0.983	0.985	0.982
	p_{50}	89	0.937	0.986	0.973	0.959	0.973	0.957	0.973	0.961
	p_{80}	36	0.860	0.938	0.938	0.901	0.936	0.913	0.937	0.903
	p_{90}	22	0.814	0.915	0.908	0.861	0.909	0.850	0.908	0.854
	p_{95}	14	0.748	0.844	0.880	0.798	0.879	0.808	0.839	0.752
Omni-MATH	p_{10}	281	0.994	0.999	0.996	0.995	0.996	0.994	0.994	0.995
	p_{20}	187	0.990	0.998	0.995	0.992	0.993	0.989	0.991	0.990
	p_{50}	74	0.968	0.993	0.985	0.977	0.982	0.973	0.984	0.982
	p_{80}	28	0.884	0.976	0.934	0.893	0.939	0.909	0.956	0.940
	p_{90}	17	0.852	0.935	0.917	0.858	0.898	0.841	0.924	0.896
	p_{95}	11	0.837	0.908	0.866	0.804	0.876	0.793	0.876	0.838

Table 17: Per-benchmark MAE (accuracy points) for MIS subsets across distance thresholds, averaged over all embedding models, distances, and solvers.

Benchmark	p_{10}	p_{20}	p_{50}	p_{80}	p_{90}	p_{95}
GPQA	3.39	4.78	6.76	10.63	12.08	16.79
IFEval	1.76	2.46	3.87	4.99	7.12	8.22
MMLU-Pro	3.14	3.75	5.59	8.47	9.95	12.36
Omni-MATH	2.95	4.65	6.71	9.77	12.14	11.48

models shift by more than 10 points, and the single worst run (MMLU-Pro, Online-MIS, `qwen4b`, cosine, p_{95}, $|S|$=3) reaches an MAE of 35.7 points (Table 18)—and panel (c) traces it to subset size, with small subsets carrying most of the error. Panel (d) confirms the errors are near-unbiased (overall mean signed shift $+0.56$ points), so subset scores neither systematically inflate nor deflate accuracy; the risk of aggressive pruning is variance, which is exactly what a practitioner controls through the threshold.

Absolute-accuracy error of the random baseline. Figure 17 repeats the MAE analysis for the size-matched random baseline, averaged over five repetitions. The random baseline behaves like MIS in every qualitative respect: its MAE rises monotonically with the threshold (panel a), follows the same benchmark ordering (GPQA hardest at 7.78 points on average; IFEval and MMLU-Pro easiest at 4.00 and 4.76), and is near-unbiased overall (mean signed shift -0.27 points), so its subset scores neither systematically inflate nor deflate accuracy. Quantitatively, however, random attains *marginally lower* MAE than MIS at every threshold (panel b; Table 19): 1.97 vs. 2.81 points at p_{10}, rising to 9.69 vs. 12.21 at p_{95}, an overall gap of roughly one to two accuracy points that widens slightly under aggressive pruning. This mirrors the small Spearman-ρ advantage of the baselines

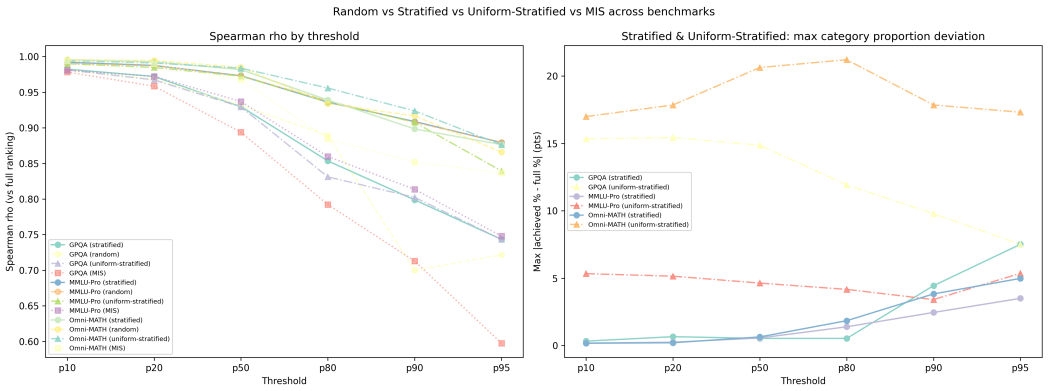


Figure 12: **Random vs. stratified vs. uniform-stratified vs. MIS, across benchmarks and thresholds.** *Left:* Spearman ρ vs. the full ranking, one line per (benchmark, method) pair. *Right:* maximum |achieved category % − full-benchmark category %| for the two stratified variants only. Uniform-stratified's deviation-from-full line sits well above stratified's at every threshold and does not close as the subset shrinks, because uniform-stratified is not targeting the full benchmark's proportions in the first place; stratified's line stays near the bottom throughout, since matching those proportions is exactly what it is built to do.

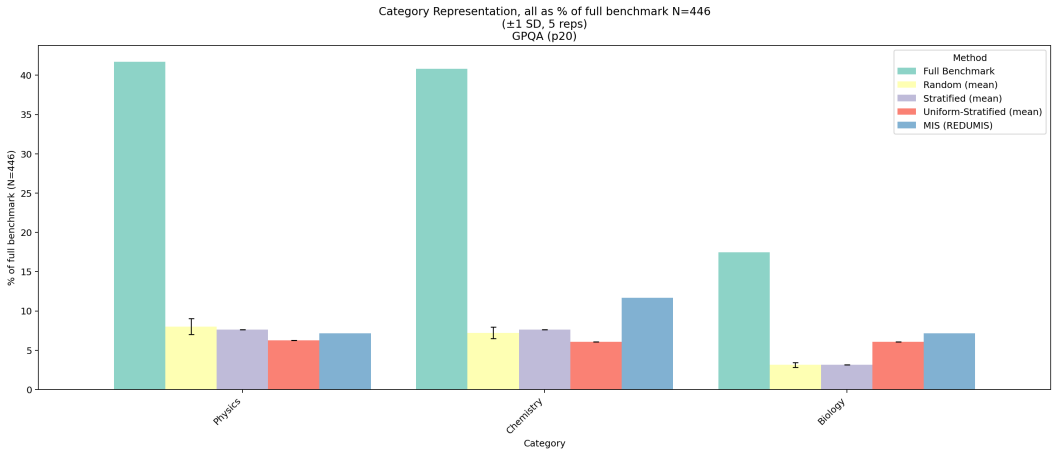


Figure 13: Category representation for GPQA (p_{20}), full benchmark vs. random, stratified, uniform-stratified (mean ± 1 SD over 5 reps, matched to the 82-prompt embedding-averaged size), and MIS (ReduMIS, `qwen8b`, standardized Euclidean, its own 116-prompt subset), all expressed as % of the full benchmark ($N = 446$).

reported in Section 5: i.i.d. sampling reproduces both the ranking and the absolute accuracy of the full benchmark slightly better on average. The distinction is therefore not accuracy fidelity but *stability*—MIS subsets change far less across seeds ($W \geq 0.90$ almost everywhere, versus baselines that drop into the seed-unstable regime under compression)—and coverage, which the size-matched random draw does not control. In short, a practitioner who only needs an unbiased absolute-accuracy estimate at a fixed budget can use random sampling, whereas MIS is preferable when reproducibility across seeds and balanced coverage matter.

Absolute-accuracy error of the stratified baseline. Figure 18 repeats the MAE analysis for the size-matched stratified baseline, which allocates the subset across a benchmark's categories *in proportion to their frequency in the full benchmark* (rather than uniformly), and is therefore defined only for the three benchmarks with category labels (GPQA, MMLU-Pro, Omni-MATH; IFEval is excluded). Its behavior matches the other selectors qualitatively: MAE rises monotonically with the threshold, keeps the same benchmark ordering (GPQA hardest at 7.47 points on average), and remains near-

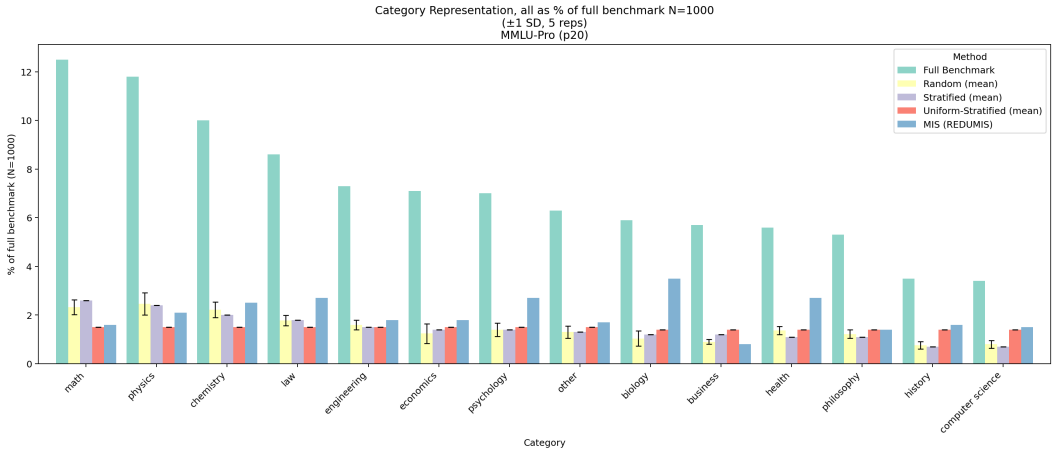


Figure 14: Category representation for MMLU-Pro (p_{20}), full benchmark vs. random, stratified, uniform-stratified (mean ± 1 SD 5 reps, matched to the 204-prompt embedding-averaged size), and MIS (ReduMIS, `qwen8b`, standardized Euclidean, its own 284-prompt subset), all expressed as % of the full benchmark ($N = 1000$).

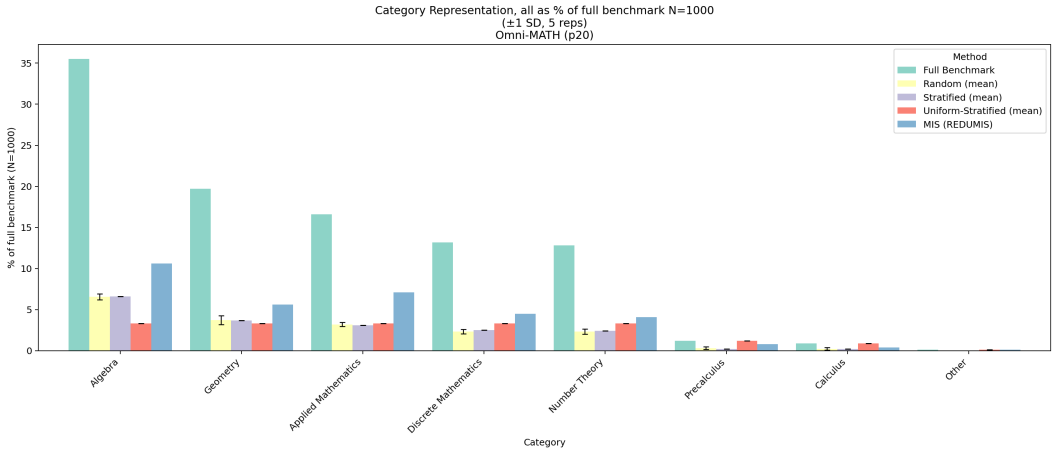


Figure 15: Category representation for Omni-MATH (p_{20}), full benchmark vs. random, stratified, uniform-stratified (mean ± 1 SD over 5 reps, matched to the 187-prompt embedding-averaged size), and MIS (ReduMIS, `qwen8b`, standardized Euclidean, its own 332-prompt subset), all expressed as % of the full benchmark ($N = 1000$).

unbiased (mean signed shift -0.23 points). Quantitatively it stays below MIS at every threshold (2.4 points vs. MIS's 2.81 at p_{10}, widening to 10.88 vs. 12.21 at p_{95}). Preserving the benchmark's category proportions thus yields no absolute-accuracy advantage over i.i.d. sampling—expected, since proportional stratification only reduces variance from category-frequency imbalance, which i.i.d. sampling already matches in expectation—consistent with its negligible effect on ranking agreement and seed stability (Section 5).

Absolute-accuracy error of the uniform-stratified baseline. Figure 19 repeats the MAE analysis for the size-matched uniform-stratified baseline, which allocates the subset *equally* across a benchmark's categories via water-filling rather than proportionally to their frequency, and like stratified sampling is defined only for the three benchmarks with category labels (GPQA, MMLU-Pro, Omni-MATH; IFEval is excluded). Its behavior is qualitatively similar to the other baselines: MAE rises monotonically with the threshold, GPQA is hardest (7.87 points on average, versus 4.73 for MMLU-Pro and 6.51 for Omni-MATH), and the errors are near-unbiased (mean signed shift $+1.13$ points, the largest of the three baselines but still small relative to the MAE itself). Quantitatively it stays below MIS at mild

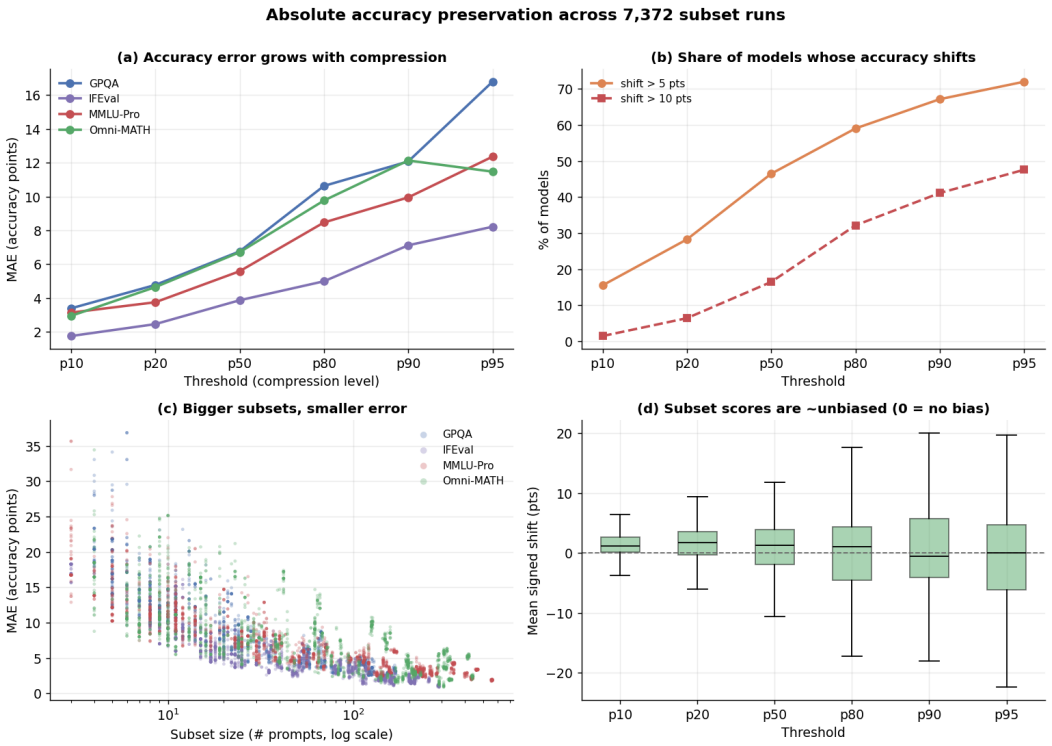


Figure 16: **Absolute accuracy preservation across** 7,372 **subset runs.** Run-level MAE is the mean over models of $|\mathrm{acc}_m^{\mathrm{subset}} - \mathrm{acc}_m^{\mathrm{full}}|$ (Eq. 5), in accuracy points. **(a)** MAE rises monotonically with the threshold (compression level) on every benchmark, from $\approx$3 points at p_{10} to 8–17 at p_{95}; IFEval is the most faithful and GPQA degrades fastest. **(b)** The share of models whose accuracy moves by more than 5 and 10 points grows from $\sim$16%/2% at p_{10} to $\sim$72%/48% at p_{95}. **(c)** MAE falls steeply with subset size (log scale): most error is incurred by very small subsets and the curve flattens once subsets reach the low hundreds. **(d)** The mean signed shift stays centered on zero at every threshold, so subsets are approximately unbiased—the growing MAE is dispersion, not directional bias.

Table 18: Worst-case subset runs by MAE. Both cases occur at the most aggressive threshold (p_{95}) with very small subsets, where a single volatile model dominates the error.

Benchmark	Solver	Embed.	Dist.	$\lvert S\rvert$	MAE	Max shift	Worst model
MMLU-Pro	Online-MIS	qwen4b	cosine	3	35.7	81.9	Kimi-K2-Instruct
GPQA	ReduMIS	qwen4b	cosine/correlation	4	20.4	72.7	Grok-4-0709

thresholds (2.23/2.75 vs. MIS's 2.81/3.91 at p_{10}/p_{20}) but the gap is noisier under aggressive pruning, spiking to 10.48 at p_{90} before easing to 10.96 at p_{95} (vs. MIS's 10.32/12.21)—a non-monotonic step driven by GPQA and Omni-MATH's few, highly imbalanced categories, where forcing equal representation leaves very few prompts per category once the subset shrinks to single digits (cf. the doubly-degraded GPQA case in Appendix F). Forcing equal rather than proportional category coverage therefore buys no consistent absolute-accuracy advantage over MIS, and its higher variance under aggressive pruning is the accuracy-side counterpart of the seed-instability result reported in Section 5.

Comparison across all baselines. Table 19 and Figure 20 bring together the four methods. All three size-matched baselines beat MIS on raw MAE at nearly every threshold, but the margin and its stability differ: random is the strongest and most consistent baseline overall, attaining the lowest MAE at five of six thresholds (losing only at p_{20}, to stratified) and needing no category information, which also makes it the only baseline defined on all four benchmarks. Stratified and uniform-stratified both trail random by a widening margin under aggressive pruning – stratified reaches 10.88 points at

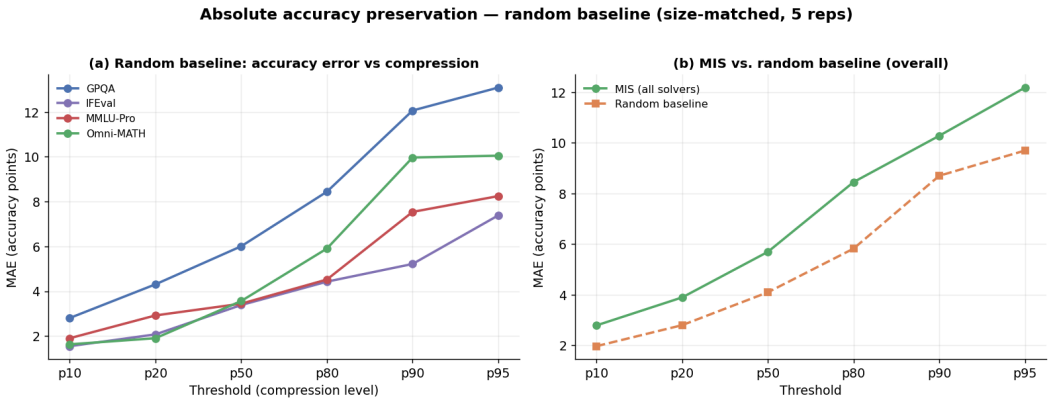


Figure 17: **Absolute accuracy preservation for the random baseline (size-matched, 5 reps). (a)** Random-baseline MAE vs. threshold per benchmark: error rises monotonically with compression, from $\sim$2 points at p_{10} to 7–13 at p_{95}, with the same benchmark ordering as MIS (GPQA hardest, IFEval/MMLU-Pro easiest). **(b)** Overall MAE of MIS (all solvers) vs. the random baseline: random attains marginally lower MAE at every threshold, mirroring its small ρ advantage; both are near-unbiased (random mean signed shift -0.27 points). Together with the seed-stability results (Sec. 5), this locates MIS's benefit in ranking *stability* and coverage rather than absolute-accuracy fidelity. See Figure 20 for the comparison against all three baselines together.

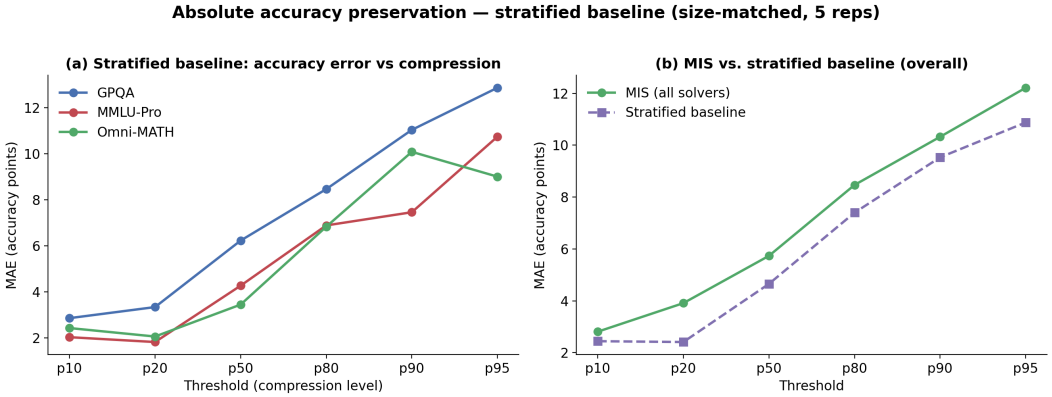


Figure 18: **Absolute accuracy preservation for the stratified baseline (size-matched, 5 reps).** Stratified sampling draws prompts so that each category keeps its *proportion* in the full benchmark (proportional allocation), and is defined only for benchmarks with category labels, so **IFEval is absent** (GPQA, MMLU-Pro, Omni-MATH only). **(a)** MAE vs. threshold per benchmark; error rises monotonically with compression, GPQA hardest. **(b)** Overall MAE of MIS (all solvers) vs. stratified: stratified stays below MIS at every threshold and is near-unbiased (mean signed shift -0.23 points). Because stratified omits IFEval, its overall average is not strictly comparable to MIS's; the per-threshold ordering is nonetheless stable. See Figure 20 for the comparison against all three baselines together.

p_{95} and uniform-stratified 10.96, versus random's 9.69—though both still stay below MIS's 12.21 at that threshold. Uniform-stratified is the least consistent of the three: comparable to the other baselines at p_{10}/p_{20}, it briefly loses its advantage over MIS entirely at p_{90} (10.48 vs. MIS's 10.32) before falling back below MIS at p_{95}, reflecting the non-monotonic behavior noted above. None of this changes the central trade-off of the paper: even where a baseline attains lower MAE than MIS, it purchases that advantage with the seed instability documented in Section 5 (Table 16), so the choice between MIS and a size-matched baseline remains one of stability and coverage versus raw absolute-accuracy fidelity, not absolute-accuracy fidelity alone.

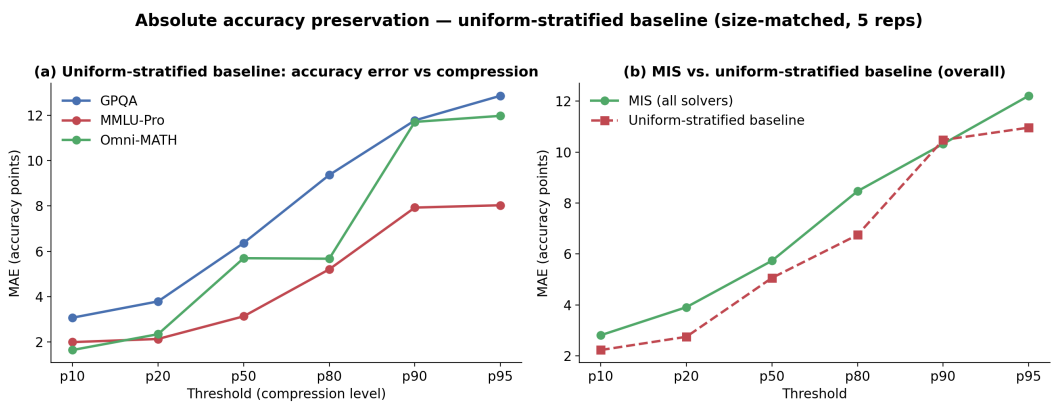


Figure 19: **Absolute accuracy preservation for the uniform-stratified baseline (size-matched, 5 reps).** Uniform-stratified sampling draws prompts so that every category receives an *equal* share of the subset via water-filling, rather than a share proportional to its frequency in the full benchmark, and like stratified sampling is defined only for benchmarks with category labels, so **IFEval is absent** (GPQA, MMLU-Pro, Omni-MATH only). **(a)** MAE vs. threshold per benchmark: error rises with compression, GPQA hardest throughout, with a pronounced late-threshold spike on GPQA and Omni-MATH once their few, imbalanced categories leave very small per-category budgets. **(b)** Overall MAE of MIS (all solvers) vs. uniform-stratified: uniform-stratified stays below MIS through p_{80} but is noisier under aggressive pruning and briefly exceeds MIS at p_{90} (10.48 vs. 10.32) before dropping back below it at p_{95}. See Figure 20 for the comparison against all three baselines together.

Table 19: Mean Absolute Error (MAE, accuracy points) across distance thresholds for MIS (averaged over all four solvers and four benchmarks), the size-matched Random baseline, and the Stratified and Uniform-Stratified baselines (GPQA, MMLU-Pro, Omni-MATH; IFEval undefined for the latter two). Lower is better.

Method	p_{10}	p_{20}	p_{50}	p_{80}	p_{90}	p_{95}
MIS (Ours)	2.81	3.91	5.73	8.47	10.32	12.21
Random Baseline	1.97	2.80	4.10	5.82	8.70	9.69
Stratified Baseline	2.44	2.41	4.65	7.40	9.53	10.88
Uniform-Stratified	2.23	2.75	5.06	6.75	10.48	10.96

H Unsupervised Structure of the Prompt Proximity Graph Feature Space

To understand what drives the topology of the complementary graphs, we perform an unsupervised clustering of all instances in the space of their graph-theoretic features and then characterize the resulting clusters *post hoc* against the generative factors of each instance. The analysis shows that the feature geometry is governed almost entirely by two factors—the edge-retention threshold and the benchmark—while the embedding size, the chunking scheme, and the embedding model leave essentially no detectable signature. The distance measure has only a second-order effect, driven by `seuclidean`; `cosine` and `correlation` are effectively interchangeable.

Data and features. Each instance is a proximity graph summarized by nine graph-theoretic features: the number of vertices and edges, density, average / maximum / minimum degree, the global clustering coefficient, the number of connected components, and the size of the largest connected component. Every instance is generated by a fully balanced factorial combination of six metadata factors, decoded from its identifier (Table 20), giving $N = 648$ instances. Two features are near-perfect proxies for design factors: the vertex count is fixed by the benchmark (GPQA$\approx$446, IFEval=541, MMLU-Pro and Omni-MATH=1000), and it correlates 0.99 with the largest-component size; density is set directly by the retention threshold. This is important for interpreting the clusters below.

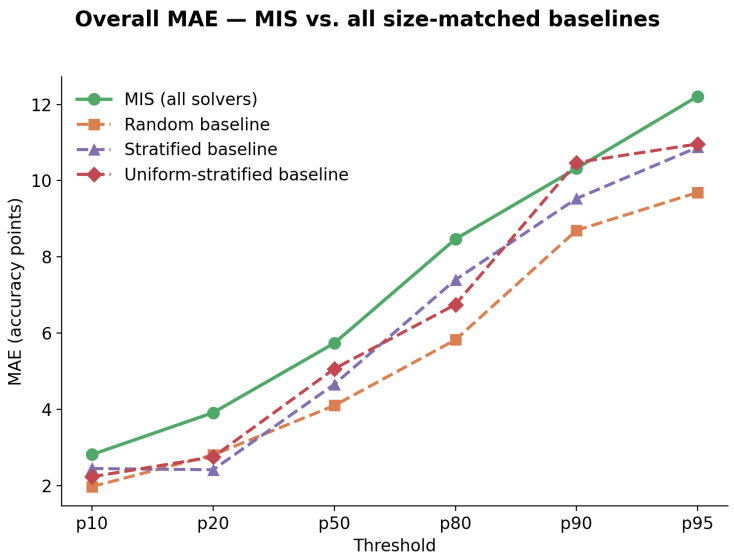


Figure 20: **Overall MAE across all four methods.** Mean absolute error (accuracy points), averaged over benchmarks, vs. distance threshold, for MIS (all solvers) and the three size-matched baselines. Random and stratified track closely and stay below MIS at every threshold; uniform-stratified matches them at mild thresholds but is noisier under aggressive pruning.

Table 20: Generative factors of each instance (balanced factorial, $N = 648$). The embedding model is nested within the embedding size.

Factor	Levels	Count/level
Embedding size	large, small	432 / 216
Chunking	chunked, nochunk (large); fixed (small)	216 each
Benchmark	GPQA, IFEval, MMLU-Pro, Omni-MATH	162 each
Embedding model	nemotron, qwen4b, qwen8b (large); bge, e5, gist (small)	144 / 72
Distance measure	correlation, cosine, seuclidean	216 each
Threshold	p10, p20, p50, p80, p90, p95	108 each

Clustering procedure. The heavy-tailed features (edges, degrees, component count, and the size features) are $\log(1+x)$-transformed; all nine features are then standardized to zero mean and unit variance. We cluster with k-means (50 restarts) in the standardized feature space. The number of clusters is selected by jointly considering the silhouette coefficient, the Davies–Bouldin index, and the agreement between k-means and Ward agglomerative clustering measured by the adjusted Rand index (ARI); see Table 21. The silhouette is maximized at $k{=}4$ (0.503), the Davies–Bouldin index is near its minimum, and k-means and Ward clustering agree *perfectly* at $k{=}4$ (ARI$=$1.000). We therefore adopt $k{=}4$; the two- and three-dimensional projections in Figure 21 are used only for visualization. Clusters are indexed by increasing mean density.

Cluster structure. The four clusters form a clean 2×2 grid over exactly two factors: a *density regime* set by the threshold and a *graph-size regime* set by the benchmark (Table 22, Figure 22). Sparse clusters (thresholds p10/p20) contain fragmented graphs with many connected components and low clustering; dense clusters (p50–p95) are near-fully connected. Within each density regime, the small-benchmark cluster (GPQA/IFEval, $\sim$450–540 vertices) separates from the large-benchmark cluster (MMLU-Pro/Omni-MATH, 1000 vertices). The sparse clusters are half the size of the dense ones (108 vs. 216) solely because the sparse regime spans two thresholds and the dense regime spans four. The projections in Figure 21 show the four groups as clearly separated in PCA, t-SNE, and UMAP, consistent with the high silhouette.

Table 21: Cluster-count selection. Silhouette (higher is better) peaks at k=4; k-means and Ward clustering agree perfectly (ARI=1) at $k \leq 4$.

k	Inertia	Silhouette ↑	Davies–Bouldin ↓	ARI (k-means vs. Ward)
2	3142.3	0.408	1.047	0.693
3	1968.7	0.460	0.873	1.000
4	**1334.2**	**0.503**	**0.775**	**1.000**
5	1053.7	0.489	0.749	0.985
6	863.1	0.470	0.822	0.936
7	748.3	0.478	0.804	0.928
8	648.3	0.480	0.806	0.902

Table 22: The four clusters as a 2×2 threshold $\times$ benchmark grid. Values are cluster means. Density and clustering are driven by the threshold; vertex count is driven by the benchmark.

Cluster	Regime	n	Density	Clust. coef.	Conn. comp.	Benchmarks / thresholds
0	sparse, large	108	0.15	0.61	70	MMLU-Pro, Omni-MATH / p10, p20
1	sparse, small	108	0.15	0.60	35	GPQA, IFEval / p10, p20
2	dense, large	216	0.79	0.89	4	MMLU-Pro, Omni-MATH / p50–p95
3	dense, small	216	0.79	0.90	4	GPQA, IFEval / p50–p95

Factor enrichment and invariance. We quantify the association between each factor and the clustering by the enrichment $\log_2(\text{observed}/\text{expected})$ of every factor level in every cluster, where the expected frequency is the level's overall frequency (Figure 23). Only *benchmark* and *threshold* deviate from zero; the embedding size, chunking scheme, and embedding model are uniformly distributed across all four clusters (enrichment $= 0$ to two decimals everywhere). Complementary-graph topology is therefore *invariant* to these three factors at the resolution of the clustering: changing the embedding size, chunking, or embedding model does not move an instance across a cluster boundary. Consistent with this, no single factor aligns strongly with the feature geometry on its own—the silhouette of every factor's own labels is small (Figure 24b; threshold 0.064, embedding 0.060, benchmark 0.046, all others ≤ 0.002)—because the observed structure is a *joint* threshold $\times$ benchmark-size partition rather than a partition along any single axis.

Effect of the distance measure. The distance measure has a real but second-order effect that lives *within* the four coarse clusters rather than forming its own top-level split (Figure 24a). Holding all other factors fixed and comparing the three distance variants of each otherwise-identical cell in the standardized feature space, `cosine` and `correlation` are almost indistinguishable (mean ℓ_2 distance 0.002; they are assigned to the same cluster in every cell), whereas `seuclidean` is a clear outlier (mean distance 0.376–0.377 from the other two). The distance measure thus perturbs the graph topology, but by an order of magnitude less than the threshold or benchmark, which is why it does not appear in the enrichment of the four clusters.

Interpretation and limitations. Because the vertex count is fixed by the benchmark and the density by the threshold, the unsupervised clustering largely *recovers two axes of the experimental design* rather than uncovering latent structure orthogonal to it. The substantive finding is therefore in what the clustering *does not* separate: the embedding size, chunking scheme, and embedding model produce complementary graphs that are statistically indistinguishable at this resolution, and among the distance measures only `seuclidean` behaves differently, with `cosine` and `correlation` interchangeable. A scale-free analysis that normalizes out the benchmark-driven size axis (e.g. clustering on density, clustering coefficient, and size-normalized component statistics) could be used to test whether any finer embedding- or distance-dependent structure emerges once the two dominant axes are removed.

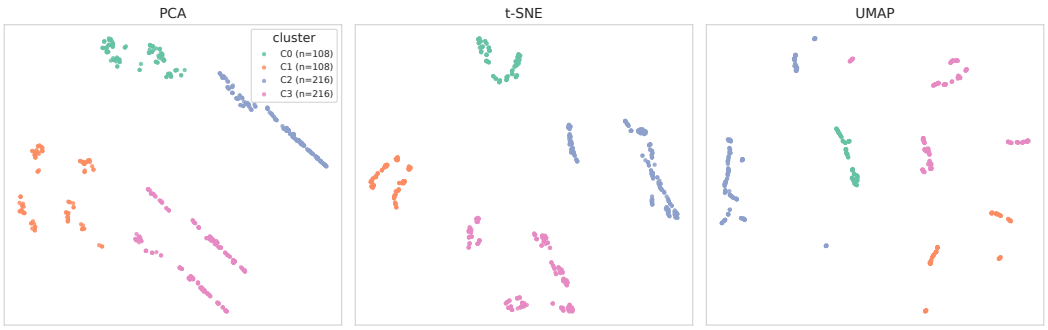


Figure 21: The four clusters in two-dimensional projections of the standardized nine-dimensional feature space (PCA, t-SNE, UMAP). Colors denote the k-means cluster; the groups are well separated in all three projections.

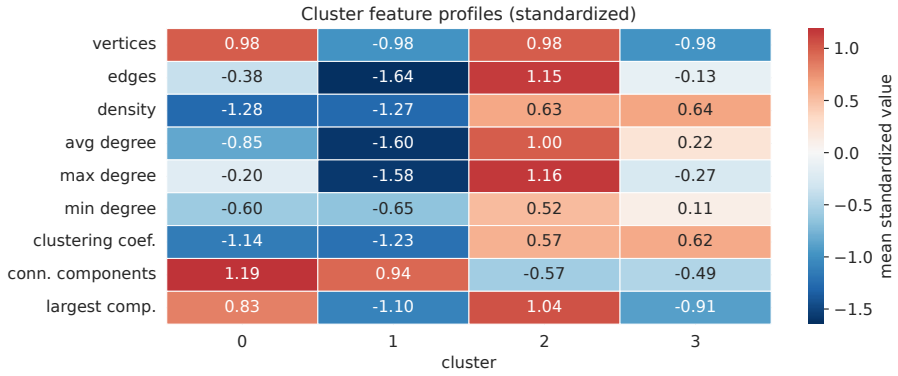


Figure 22: Standardized feature profile of each cluster (mean z-score per feature). Clusters 0/1 (sparse) have low density, low clustering, and many components; clusters 2/3 (dense) are the opposite. Within each density regime, clusters differ on the size features (vertices, largest component), i.e. by benchmark.

I Distance Distribution Plots

Figures 25–27 show pairwise-distance distributions for GPQA under different embedding models and distance measures, motivating percentile-based thresholding over fixed absolute values.

J Rank-Shift Heatmap

Figure 28 presents the rank-shift heatmap for the case study of Section 5 (Online-MIS, IFEval, `qwen4b`, cosine, p_{20}). Each cell encodes the deviation of a model's subset rank from its full-benchmark rank (green = higher on the subset, red = lower); the rightmost column is the full baseline (zero by definition). The seed columns show near-identical patterns ($W = 0.994$), while a few models shift relative to the baseline, giving $\rho = 0.968$: the subset is highly stable across seeds yet re-weights a small number of models whose standing depends on the pruned region.

K Embedding Similarity vs. Behavioral Similarity

Figure 29 tests whether, within the retained–discarded pairs of Section 5 (`qwen8b`, seuclidean, p_{20}), a smaller embedding distance predicts more similar model behavior. For each pair we plot the embedding distance against the cosine similarity of the two prompts' 66-LLM response vectors. Because every prompt is paired with its nearest discarded neighbor, distances occupy a narrow band whose location is set per benchmark by the p_{20} threshold: GPQA and MMLU-Pro concentrate near

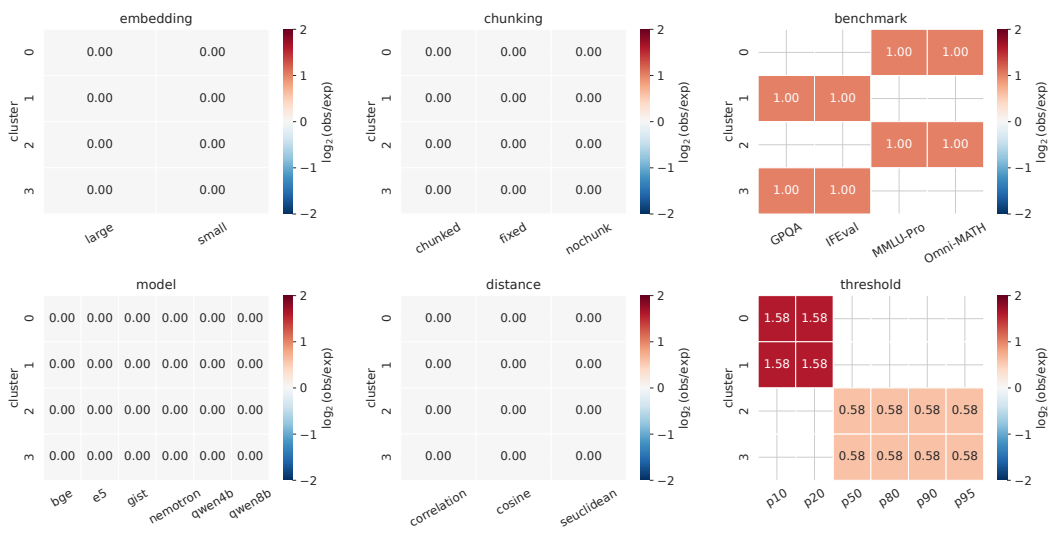


Figure 23: Post-hoc factor enrichment per cluster, $\log_2(\text{obs}/\text{exp})$. Red indicates over-representation, blue under-representation, and white indicates the level is absent from that cluster. Only *benchmark* and *threshold* discriminate the clusters; *embedding*, *chunking*, and *model* are flat at zero, indicating invariance.

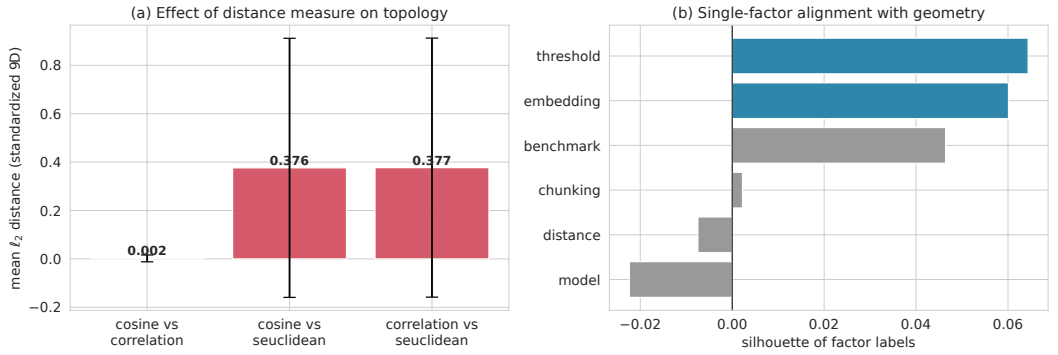
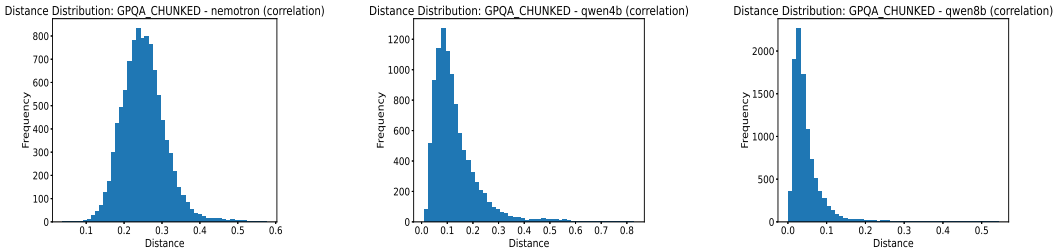


Figure 24: (a) Mean ℓ_2 separation between distance-measure variants of the same cell in standardized feature space: `cosine` and `correlation` are interchangeable, while `seuclidean` diverges. Error bars are standard deviations over cells. (b) Silhouette of each factor's own labels in the feature space; all values are small, indicating that no single factor explains the geometry alone.

0.16, Omni-MATH near 0.25, reflecting its more dispersed embedding geometry. The benchmarks instead differ chiefly on the y-axis, tracking difficulty rather than geometry: MMLU-Pro (easiest, 68% mean solve rate) sits in a high, compact cosine ridge; GPQA (mid-difficulty) is centred lower; and Omni-MATH (hardest, 40% solve rate) spreads across the full cosine range. Slopes are flat in all three ($|r| \leq 0.10$), so within the threshold band the exact distance barely predicts behavioral similarity.

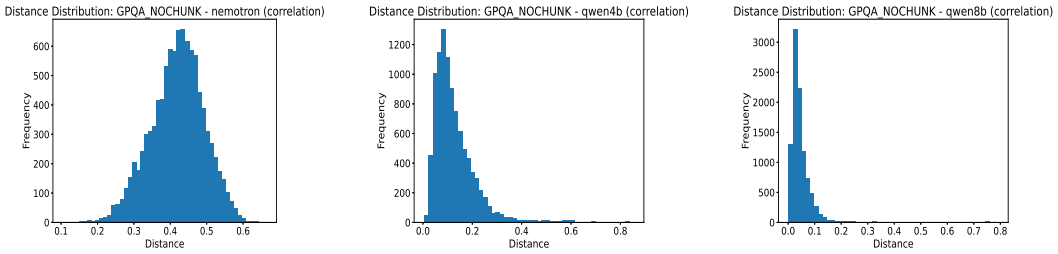
L Computational Details

Large embedding generation. Dense representations were generated with Qwen3-8B, Qwen3-4B, and Nemotron-8B in float16 via HuggingFace Transformers. Each prompt was embedded under two conditions. *No-chunk:* hard truncation at 4,096 tokens, single forward pass, mean pooling over non-padding positions, L2 normalization. *Chunked:* overlapping windows of 510 tokens (stride 256, 50% overlap); each chunk embedded independently and combined as a length-weighted average, then L2-normalized. For Nemotron-8B each prompt was prefixed with "`Instruct: Represent`



(a) Nemotron (b) Qwen4B (c) Qwen8B

Figure 25: Distance distributions for GPQA_CHUNKED across models (correlation distance).



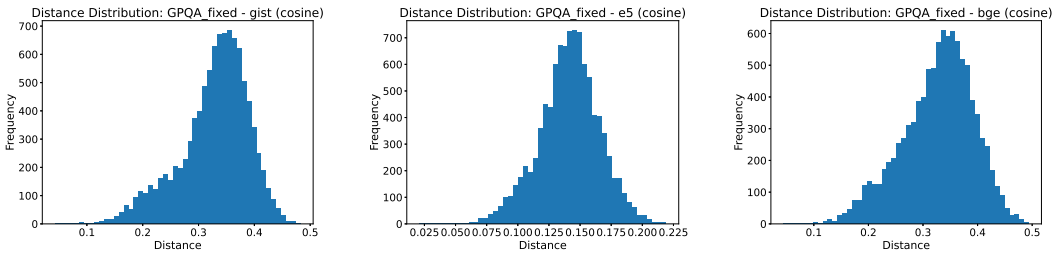
(a) Nemotron (b) Qwen4B (c) Qwen8B

Figure 26: Distance distributions for GPQA_NOCHUNK across models (correlation distance).

`the following text for clustering:` "; no prefix was used for Qwen models. The full pipeline over all four datasets, both conditions, and all three models completed in 52 min 42 s of wall-clock time on a node with 2× NVIDIA L40S GPUs, a 64-core AMD CPU, and ∼503 GB RAM (SLING HPC).

Small embedding generation. We additionally used three dedicated sentence encoders: BGE-large-en-v1.5, E5-large, and GIST-small-Embedding-v0, run locally (Intel Core Ultra 7 258V, 32 GB RAM, Intel Arc 140V, 16 GB VRAM). Task-specific prefixes were used per model (BGE: "`Represent this sentence for searching relevant passages:` "; E5: "`query:` "; GIST: none). Non-overlapping 510-token chunks were used given the 512-token context limit, with length-weighted aggregation and L2 normalization.

Optimization experiments. MIS experiments ran on the VEGA EuroHPC system (Atos BullSequana XH2000; dual-socket AMD EPYC 7H12 @ 2.6 GHz; 256 GB–1 TB RAM; HDR100 InfiniBand; SLURM). All MIS runs were single-threaded for fair comparison, with 32 GB RAM and a 600 s time limit per job; stochastic solvers used five seeds.



(a) GIST (b) E5 (c) BGE

Figure 27: Distance distributions for GPQA across the small encoder models (cosine distance).

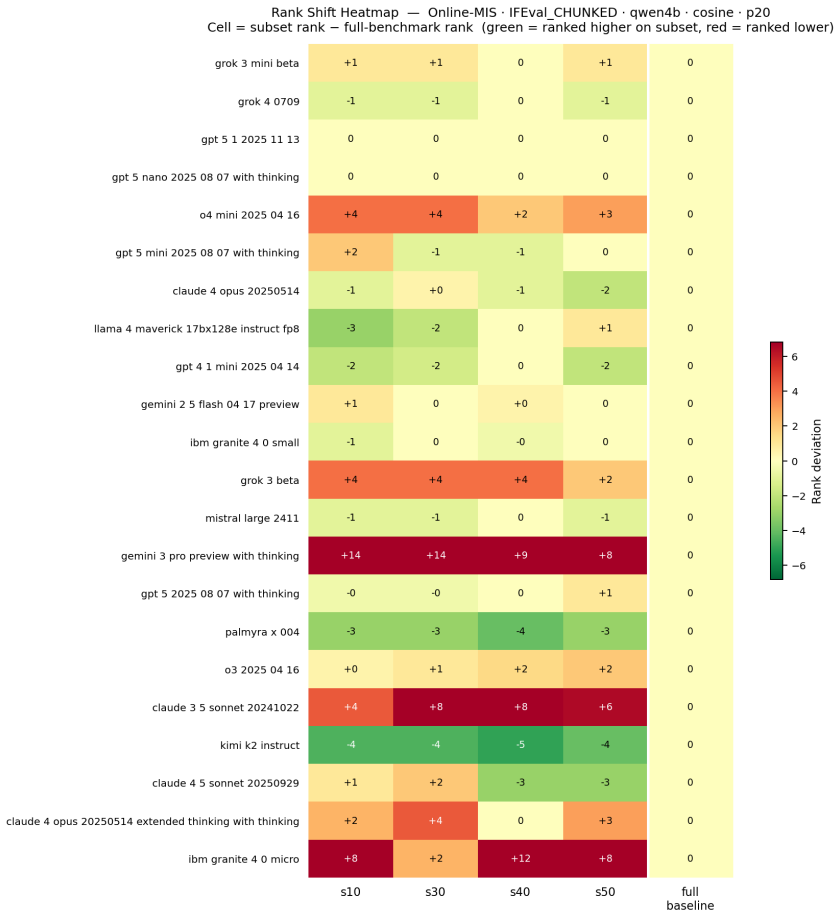


Figure 28: Rank-shift heatmap for Online-MIS on IFEval (`qwen4b`, cosine, p_{20}). Columns $s10$–$s50$ are seeds; the last column is the full-benchmark baseline. Seeds agree closely ($W = 0.994$) while the ranking differs from the baseline on a few models ($\rho = 0.968$).

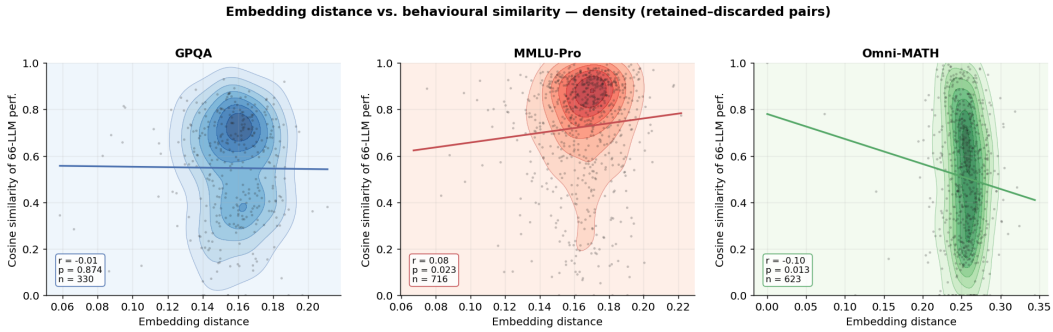


Figure 29: **Embedding distance vs. behavioural similarity, by benchmark.** Embedding distance (x) against cosine similarity of the two prompts' 66-LLM response vectors (y) for each retained–discarded pair (`qwen8b`, seuclidean, p_{20}). Contours: 2D kernel-density estimate; grey points: individual pairs; line: linear fit with Pearson r.