The Flaw of Averages: Measuring Benchmark-Level Distributional Robustness

Arda Uzunoğlu and Tianjian Li and Daniel Khashabi
Department of Computer Science, Johns Hopkins University
{auzunog1,tli104,dkhasha1}@jhu.edu

Abstract

Benchmarks are central to measuring progress in language models, but aggregate scores can obscure substantial variation across subdomains, making models appear broadly competent despite concentrated strengths and weaknesses. We study this issue as *benchmark-level distributional robustness*: whether aggregate scores faithfully reflect performance across benchmark subdomains. We operationalize this notion with benchmark HARMONY, an entropy-based measure of how uniformly model performance is distributed across subdomains. Measuring HARMONY on 19 language model benchmarks across five model families, we find substantial variation in benchmark-level distributional robustness. Low-HARMONY benchmarks are more likely to yield aggregate scores that overstate broad competence, whereas high-HARMONY benchmarks provide more representative summaries of model capability. Rebalancing benchmarks by pruning overrepresented subdomains to increase HARMONY substantially shifts aggregate scores for low-HARMONY benchmarks, but leaves high-HARMONY benchmarks comparatively stable. For example, while BoolQ remains comparatively stable as HARMONY increases, PubMedQA, which evaluates performance in a medically consequential domain, exhibits substantial, often statistically significant, shifts in aggregate accuracy. Together, these findings show that aggregate scores can misrepresent broad competence when performance is unevenly distributed. We therefore recommend reporting benchmark HARMONY alongside aggregate accuracy as a diagnostic of benchmark representativeness when interpreting claims about broad model competence.

1 Introduction

Benchmarks are central to measuring and shaping scientific progress in language models, forming a feedback loop with model development (Hardy et al., 2024; Yao, 2025). As benchmarks guide model design, stronger models expose benchmark limitations. In this reciprocal process, it is essential that benchmark evaluation remains representative: aggregate scores should reflect genuine capabilities rather than evaluation artifacts. However, whether these scores are actually representative of broad competence (Swayamdipta et al., 2020; Sainz et al., 2023) has received less attention than the algorithmic advances that benchmarks are used to evaluate (Brown et al., 2020; Ouyang et al., 2022; DeepSeek-AI et al., 2025).

Motivated by this gap, we study evaluation through the lens of what we call *benchmark-level distributional robustness*: whether an aggregate score remains a representative summary of performance across the subdomains a benchmark is intended to cover. This perspective complements prior work on benchmark robustness, which has identified concerns including redundancy (Polo et al., 2024; Perlitz et al., 2024a), uneven data distributions (Huang et al., 2025), and whether benchmark gains are trustworthy (Singh et al., 2025) or support the claims commonly drawn from them (Ruan et al., 2024; Heineman et al., 2025). More specifically, we isolate a failure mode in which aggregate benchmark scores misrepresent broad competence by obscuring concentrated strengths and weaknesses. This perspective is internal to the benchmark itself, complementing both prior work on robustness to distribution shift and worst-group failure (Sagawa et al., 2020; Koh et al., 2021) and efforts to assess benchmark design more systematically (Reuel et al., 2024; Gebru et al., 2021).

To study *benchmark-level distributional robustness*, we introduce benchmark HARMONY, a benchmark-level diagnostic that measures how uniformly a model's performance is distributed across a benchmark's subdomains. As illustrated in Figure 1, we partition each benchmark into semantic subdomains, measure model performance on

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 43186–43226
October 24-29, 2026 ©2026 Association for Computational Linguistics

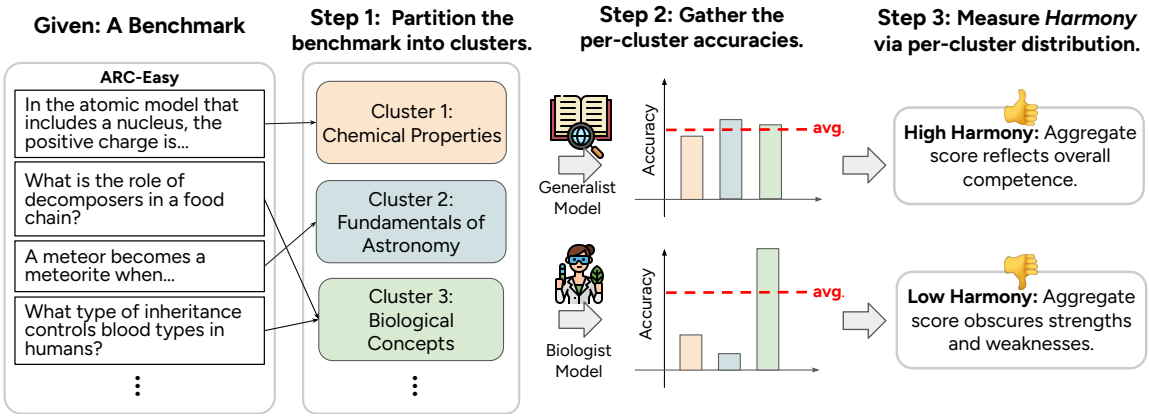


Figure 1: **Pipeline of evaluating HARMONY for a given benchmark.** Step 1: We partition the benchmark into semantic clusters (subdomains). Step 2: We gather each model's performance on every cluster. Step 3: We compute HARMONY, the uniformity of the distribution of performance across subdomains. High HARMONY indicates that aggregate benchmark scores more faithfully summarize competence across subdomains, whereas low HARMONY indicates that aggregate scores may be dominated by only part of the benchmark.

each, and compute HARMONY for each benchmark-model pair. While subdomain-level breakdowns are informative, they do not by themselves quantify whether an aggregate score is representative across subdomains or enable compact comparison across benchmarks and models. High HARMONY indicates that aggregate scores are more representative of broad competence across subdomains, whereas low HARMONY suggests that aggregate scores may be driven disproportionately by a subset of subdomains. Using HARMONY, we analyze a range of benchmarks and models to assess benchmark-level distributional robustness (§3, §4). We recommend reporting benchmark HARMONY alongside aggregate accuracy when interpreting benchmark results, as a diagnostic of whether the benchmark's score is representative across subdomains.

Our contributions are:

- **A benchmark-centric perspective on distributional robustness.** We formalize benchmark-level distributional robustness and show that aggregate benchmark scores can systematically misrepresent broad competence when performance is uneven across subdomains.
- **A diagnostic metric of benchmark representativeness across subdomains.** We introduce HARMONY, an entropy-based measure of how uniformly performance is distributed across a benchmark's subdomains, and argue that it should be reported alongside aggregate accuracy as a diagnostic of benchmark representativeness.
- **A large-scale empirical map.** We map 19 MCQA benchmarks and MMLU subtasks into a cross-model HARMONY mean-variance plane across 36 models from five families.

2 Benchmark HARMONY

2.1 Preliminaries and Notation

Let $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$ be a benchmark consisting of N input-output pairs (x, y). Our goal is to understand the *underlying distribution* of $\mathcal{B}$ by inducing a semantic partition $\mathcal{G} = \{A_1, \ldots, A_k\}$ of $\mathcal{B}$, where $A_i \subseteq \mathcal{B}$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\bigcup_{i=1}^{k} A_i = \mathcal{B}$. The partition is guided by a similarity function $\mathcal{S} : \mathcal{X} \times \mathcal{X} \to (0, 1]$ that measures the semantic similarity between data points $x_i, x_j \sim \mathcal{B}$. Lastly, let f be a model and let $\Psi(f; A_i)$ denote a measure of performance (e.g., accuracy, pass rate, or judge score) for f computed on a subdomain $A_i \subseteq \mathcal{B}$.

2.2 HARMONY: A Measure of Performance Uniformity

Intuition. Consider a biology benchmark spanning microbiology, animal biology, and plant biology. If one subdomain dominates and a model performs well only there, the aggregate score may misleadingly suggest broad biological competence. Even when subdomains are balanced, large performance differences across them can still make the average uninformative. Our goal is therefore to quantify, for each model, whether benchmark performance is broadly distributed across subdomains or concentrated in only part of the benchmark. We capture this with HARMONY, a measure of the uniformity of subdomain-level performance. One

could instead report the full vector of subdomain accuracies, which is useful descriptively, but that vector does not by itself provide a single measure of whether the aggregate score is representative or support straightforward comparison. Therefore, HARMONY is intended as a compact diagnostic. Since HARMONY depends jointly on subdomain prevalence and model-specific variation, it is not merely a measure of dataset balance. Instead, it asks whether the aggregate score is representative for that model: high HARMONY means that aggregate accuracy reflects subdomain-level performance well, whereas low HARMONY means that it does not, whether because of benchmark imbalance (e.g., overrepresented subdomains) or genuine heterogeneity in model capability across tasks. By aggregating HARMONY across models, we then assess whether a benchmark reliably supports claims of broad competence.

Formal definition of HARMONY. Given a partition $\mathcal{G}_f = \{A_i\}_{i=1}^k$, HARMONY measures how uniformly performance is distributed across the subdomains in this partition. Unlike dispersion statistics such as the variance of subdomain accuracies, which quantify the magnitude of variation across subdomains, HARMONY is designed to characterize how performance is distributed across them. For each subdomain A_i, let $w_i = \frac{|A_i|}{|\mathcal{B}|}$ denote its relative size and let $\mu = \sum_{i=1}^k w_i \Psi(f; A_i)$ be the weighted mean performance. We convert deviations of $\Psi(f; A_i)$ from μ into smooth proximity scores via a Gaussian kernel:

$$K_i \;=\; \exp\Big(- \big(\tfrac{\Psi(f;A_i)-\mu}{b}\big)^2 \Big),$$

where $b > 0$ is a bandwidth parameter.[1] We then define *performance masses*

$$p_i \;=\; \frac{w_i K_i}{\sum_{j=1}^k w_j K_j}, \qquad \sum_{i=1}^k p_i = 1,$$

and compute HARMONY as the normalized Shannon entropy

$$H(\mathcal{G}_f) \;=\; -\frac{1}{\log k} \sum_{i=1}^k p_i \log\big(p_i + \varepsilon\big) \in [0,1],$$

with a small $\varepsilon = 10^{-12}$ for numerical stability. Equivalently,

$$p_i \propto w_i \exp\left(- \left(\frac{\Psi(f; A_i) - \mu}{b} \right)^2 \right),$$

which is a weighted softmax over negative squared deviations from the benchmark mean. Subdomains whose performance lies far from the benchmark mean receive exponentially smaller mass, which lowers entropy. Thus, HARMONY is high when performance is broadly and evenly distributed across subdomains, and low when performance is concentrated in only a few of them. We defer the further discussion of HARMONY design choices to App. C.

Interpreting HARMONY. Let Π be a partitioning rule that maps a benchmark-model pair $(\mathcal{B}, f)$ to a partition $\mathcal{G}_f(\mathcal{B}) = \Pi(\mathcal{B}; f)$ using $\mathcal{S}$. We define the per-model HARMONY of $\mathcal{B}$ as

$$H_{\mathcal{B}}(f) \;:=\; H\big(\mathcal{G}_f(\mathcal{B})\big) \in [0,1].$$

Given a model set $\mathcal{F}$, we summarize $\mathcal{B}$ by the cross-model mean and variance

$$\begin{aligned} \mu_H(\mathcal{B}) &= \mathbb{E}_{f\sim\mathcal{F}}[H_{\mathcal{B}}(f)]\,, \\ \sigma_H^2(\mathcal{B}) &= \mathrm{Var}_{f\sim\mathcal{F}}(H_{\mathcal{B}}(f))\,. \end{aligned} \tag{1}$$

Since the induced partition depends on the model, benchmark-level distributional robustness is a property of a benchmark relative to a model set $\mathcal{F}$. In particular, higher $\mu_H(\mathcal{B})$ means that aggregate performance is, on average, more representative of subdomain-level performance, while lower $\sigma_H^2(\mathcal{B})$ means that this representativeness is more consistent across models. We therefore interpret benchmarks with high mean HARMONY and low variance as more distributionally robust, in the sense that their aggregate scores more reliably summarize broad competence across subdomains (rather than robustness to external shift). However, this notion is comparative rather than absolute: all else equal, $\mathcal{B}_1$ is preferred to $\mathcal{B}_2$ when it achieves higher mean HARMONY and lower variance. Accordingly, HARMONY should not be interpreted as a measure of benchmark quality overall, but as a diagnostic of one specific aspect of it (i.e. representativeness). Lastly, we note that we use model-specific partitions because benchmark-level distributional robustness concerns whether a benchmark's aggregate score is representative for a particular model, which depends on the subdomain structure relevant

[1] We set b by a robust scale of $\{\Psi(f; A_i)\}$. Let $\tilde{a} = \mathrm{median}_i\, \Psi(f; A_i)$ and $\mathrm{MAD} = \mathrm{median}_i |\Psi(f; A_i) - \tilde{a}|$, then $b = \max\{0.02,\ 1.4826 \cdot \mathrm{MAD}\}$.

to that model's behavior. A fixed human-defined partition instead measures uniformity with respect to a prespecified ontology, which may be unavailable, incomplete, or too coarse. By inducing partitions from predictive similarity, we recover the subdomains that govern a model's performance on the benchmark, so that HARMONY evaluates whether aggregate scores are representative of model behavior rather than of an externally imposed taxonomy.

2.3 Inducing Benchmark Subdomains

To compute HARMONY, we require a semantic partition of the benchmark into subdomains. Because such subdomains are typically not given a priori, we infer them from model-aware semantic similarity. Specifically, we define **predictive similarity**, a similarity measure between data points based on the divergence of their predictive distributions under a model, and induce $\mathcal{G}_f$ via spectral clustering on the resulting affinity matrix $\mathcal{S}$.

Predictive Similarity. We define predictive similarity to measure how similarly a model f behaves on two inputs at the level of its distributions. For $x_i, x_j \sim \mathcal{B}$, the predictive similarity is

$$S(x_i, x_j) = \exp\Bigg(-\tfrac{\tau}{2}\Big[D_{\mathrm{KL}}\big(\bar{p}_f(x_i) \,\|\, \bar{p}_f(x_j)\big) + D_{\mathrm{KL}}\big(\bar{p}_f(x_j) \,\|\, \bar{p}_f(x_i)\big)\Big]\Bigg). \quad (2)$$

where $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ denotes the KL divergence, τ is a scaling factor chosen as the reciprocal of the median symmetric divergence, and the averaged predictive distribution is given by $\bar{p}_f(x_i) = \frac{1}{T}\sum_{t=1}^{T} p_f(x_i, y_i^{<t})$, with $y_i^{<t}$ denoting the ground-truth prefix up to token $t-1$.[2] We defer further discussion of predictive similarity to App. B.

Clustering. Given the predictive similarity matrix $S \in (0,1]^{N\times N}$, we induce the partition of a benchmark via spectral clustering (Ng et al., 2001). We treat S as a precomputed affinity, form the symmetric normalized Laplacian $L = D^{-1/2}(D - S)D^{-1/2}$ with $D = \mathrm{diag}(S\mathbf{1})$, compute the k eigenvectors of L associated with its smallest eigenvalues, and apply k-means in this spectral embedding to obtain a partition $\mathcal{G} = \{A_1, \ldots, A_k\}$. To choose the number of subsets, we sweep $2 \le k \le 20$ and select the value maximizing the silhouette score $s(k) \in [-1, 1]$ as a data-driven measure of cluster compactness and separation (Rousseeuw, 1987). The selected k is not significantly associated with HARMONY ($r = 0.274$, $p = 0.282$; App. H). The induced subdomains are human interpretable, as we assign concise labels to clusters using a bottom-up labeling procedure and provide representative examples in Apps. P and Q.

Although predictive similarity requires access to model predictive distributions, this is only needed for model-specific partition induction. Given any fixed partition, including model-agnostic partitions induced from embeddings, HARMONY requires only subdomain level performance and is therefore applicable to black-box or proprietary models. We validate this setting in App. D.

[2]For $t > 1$, we condition on the ground-truth answer tokens rather than on the model's own autoregressive predictions, ensuring that accumulated model errors do not affect the similarity measure.

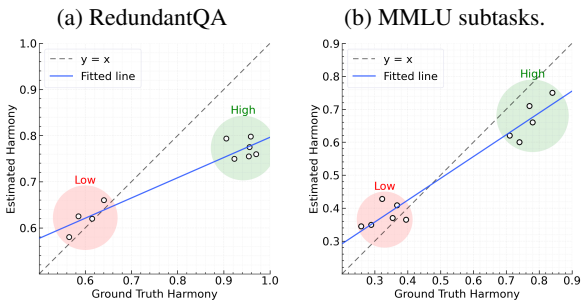


Figure 2: **Validation of our approach on (a) RedundantQA and (b) MMLU subtasks.** Estimated HARMONY strongly correlates with the ground truth and clearly separates low from high HARMONY variants. Each dot represents one variant averaged across five random seeds. Full results are provided in Fig. 7 and 8.

2.4 Validating the Induced Subdomains

Because HARMONY depends on the induced subdomain partition, validating the partitioning step is essential. To this end, we introduce RedundantQA, a synthetic four-domain[3] MCQA benchmark with a ground-truth underlying structure. Each item pairs a reference question with two *true-similar* paraphrases (same underlying knowledge) and two *false-similar* distractors (high lexical overlap, different answers). This design cleanly separates semantic from lexical similarity while allowing controlled variation in the underlying data distribution. See Appendix A for construction and validation details, along with representative examples. We validate both the induced partitions and the result-

[3]Biology, History, Economics, Popular Culture.

ing HARMONY estimates on controlled variants of RedundantQA and on a compilation of MMLU high school subtasks. As shown in Fig. 2, HARMONY estimated by our method exhibits a strong positive correlation with ground-truth HARMONY and clearly separates low- and high-HARMONY regimes. We further validate predictive similarity along three axes and defer the details to Appendix B: (i) discrimination of semantic vs. lexical similarity (App. B.2), (ii) recovery of ground-truth domains on RedundantQA and MMLU (App. B.3), and (iii) fidelity of HARMONY estimates under distributional shifts (App. B.4).

3 Main Analysis: How Harmonious Are the Benchmarks?

We now turn to the central question we investigate: how robustly do benchmarks evaluate model competence across subdomains? To answer this, we compute HARMONY for each benchmark-model pair as described in §2, and then aggregate these scores across models to place each benchmark in the mean-variance plane defined by $(\mu_H(\mathcal{B}), \sigma_H^2(\mathcal{B}))$ in Eq. 1. This representation lets us compare benchmarks by both their average distributional robustness and the stability of that robustness across models. We describe the experimental setup (§3.1) and then interpret the resulting benchmark landscape (§3.2).

3.1 Experimental Setup

We conduct evaluations using `lm-evaluation-harness`[4] across a wide range of model sizes from five families: `Llama 3` (Grattafiori and et al, 2024), `Qwen3` (Yang et al., 2025), `Gemma 3` (Team et al., 2025), `Phi-3` (Abdin et al., 2024), and `OLMo 2` (OLMo et al., 2025) (see App. M for the full list). Our setup spans 19 MCQA benchmarks covering a broad range of capabilities:

- **Reasoning**: ARC (Clark et al., 2018), ART (Bhagavatula et al., 2020), BoolQ (Clark et al., 2019), CommonsenseQA (Talmor et al., 2018), COPA (Roemmele et al., 2011), LogiQA (Liu et al., 2020), PIQA (Bisk et al., 2020), QUARTZ (Tafjord et al., 2019), SocialIQA (Sap et al., 2019a), StrategyQA (Geva et al., 2021).
- **Math Problem-Solving**: AQUA-RAT (Ling et al., 2017), MathQA (Amini et al., 2019).
- **World Knowledge**: GPQA (Rein et al., 2023), MMLU (Hendrycks et al., 2020), OpenBookQA (Mihaylov et al., 2018a), PubMedQA (Jin et al., 2019), SciQ (Welbl et al., 2017).
- **Truthfulness**: TruthfulQA (Lin et al., 2021).

For evaluation, we use average token-level log-likelihood over answer options, selecting the option with the highest average log-probability. We follow each benchmark's evaluation protocol as implemented in the harness, use zero-shot evaluation by default, and report accuracy. We focus on MCQA benchmarks because (i) the discrete label space $\mathcal{Y}$ yields unambiguous ground truth and exact accuracy,[5] and (ii) evaluation is automatic and does not rely on a judge, thereby avoiding the grading variance common in free-form scoring (Chiang et al., 2024). At the same time, HARMONY is not limited to MCQA: App. E demonstrates its application to open-ended mathematical reasoning and instruction following evaluation.

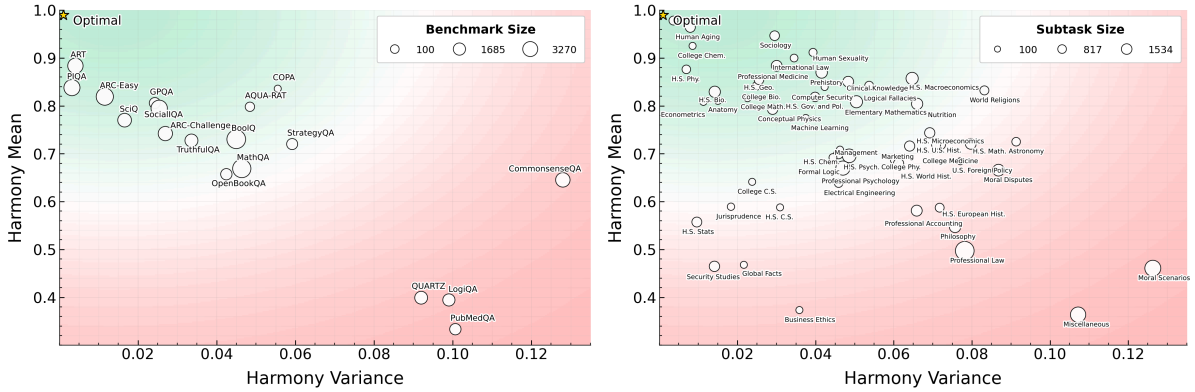
3.2 Benchmark Landscape in the HARMONY Mean-Variance Plane

We analyze the benchmarks with HARMONY by inducing model-specific partitions, computing a HARMONY score for each benchmark-model pair, and then placing each benchmark in a two-dimensional plane whose axes are the mean and variance of these per-model scores.

Fig. 3a and Fig. 3b position, respectively, the benchmark suite and the MMLU subtasks in the cross-model mean-variance plane of HARMONY. The vertical axis is $\mu_H(\mathcal{B})$, the average uniformity of performance across subdomains, and the horizontal axis is $\sigma_H^2(\mathcal{B})$, the cross-model variability of that property. Benchmarks in the *upper-left* region (high mean, low variance) tend to provide more representative and more stable aggregate evaluations across models. By contrast, low mean indicates that performance is skewed across subdomains on average, while high variance indicates that this robustness is model-dependent. Thus, upward and leftward directions in the plane correspond to more robust evaluation, whereas downward and rightward movement signals greater concentration of performance on a few subdomains and less stable conclusions across models. See App. G, I, and H for a confounder analysis covering benchmark size, pruning-rate sensitivity, and dependence on the

[4] https://github.com/EleutherAI/lm-evaluation-harness

[5] In free-response settings, multiple plausible intermediate steps can lead to the same answer, complicating evaluation.



(a) MCQA Benchmarks.

(b) MMLU Subtasks

Figure 3: **Mean-variance plane for HARMONY across (a) MCQA benchmarks and (b) MMLU subtasks.** Each point represents a benchmark or a subtask plotted by the HARMONY mean ($\mu_H(\mathcal{B})$) and variance ($\sigma_H^2(\mathcal{B})$) over 36 models. Upper-left (high mean, low variance) indicates higher *benchmark-level distributional robustness*; rightward (higher variance) and downward (lower mean) shifts signal diminished robustness. In (a), HARMONY mean and variance show a strong negative Pearson correlation ($r = -0.785$ with $p = 0.0001$), while in (b) the correlation is more moderate ($r = -0.404$ with $p = 0.0025$).

number of clusters k. App. D further shows that the MCQA benchmark landscape is qualitatively and quantitatively similar when using fixed model-agnostic MiniLM partitions.

4 Further Analyses

In this section, we pursue two controlled analyses. First, we show how less harmonious benchmarks can distort model evaluation by making aggregate scores less representative of performance across subdomains. Second, we examine whether this fragility is especially pronounced for larger models or for models trained on more tokens.

4.1 Effects of Benchmark Rebalancing

We now study whether less harmonious benchmarks are more vulnerable to unrepresentative aggregate evaluation. To do so, we prune benchmarks using predictive similarity to remove overly similar items. The pruning ratio is set inversely proportional to benchmark HARMONY, so that high-HARMONY benchmarks receive minimal pruning while low-HARMONY benchmarks are pruned more aggressively.[6] This procedure reduces benchmark skewness and lets us examine how aggregate accuracy changes under a rebalancing intervention that typically increases HARMONY.

As shown in Fig. 4, accuracies on high-HARMONY benchmarks remain stable under pruning, with changes that are not statistically significant despite increased HARMONY. By contrast, low-HARMONY benchmarks are fragile: pruning substantially increases HARMONY and is accompanied by statistically significant accuracy shifts (see App. J). Under this rebalancing intervention, per-subdomain accuracies often tighten around the benchmark mean, making the aggregate score a more representative summary of performance across subdomains. Overall, these results suggest that low-HARMONY benchmarks are more vulnerable to aggregate evaluations that are less representative of broad competence, whereas high-HARMONY benchmarks tend to be more stable under pruning. While Fig. 4 shows this pattern for the Qwen3 family, App. N provides the full results for other model families.

4.2 Effects of Model Scale and Token Budget

We next examine how HARMONY varies with model size and pre-training token budget.

Model parameters. We find that the relationship between model parameter count and HARMONY is **family-specific**. As shown in Fig. 5, within-family comparisons reveal a negative correlation for the Qwen and Llama families, indicating that larger models in these families concentrate performance more on a few subdomains. In contrast, the Gemma and OLMo families exhibit a positive correlation between model size and HARMONY, with larger models distributing performance more evenly across benchmark subdomains. Parameter count alone is therefore not a reliable predictor of

[6] Specifically, we use the formula $p = \mathrm{clip}_{[0.05,0.5]}\big(0.05 + (0.5 - 0.05)\big(\frac{1-\mathrm{clip}(H;0.1,1)}{1-0.1}\big)^{1.5}\big)$.

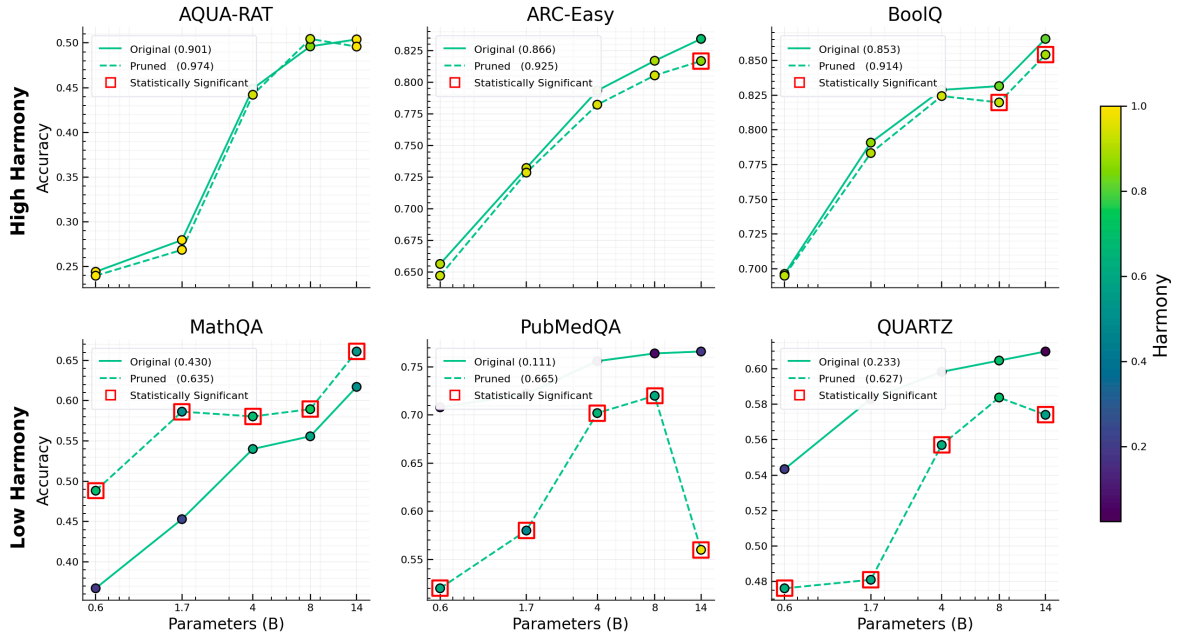


Figure 4: **Balancing benchmarks via pruning.** We remove overly similar items. The *top row* shows more harmonious benchmarks, where accuracy remains stable as HARMONY increases. The *bottom row* shows less harmonious benchmarks, where HARMONY rises and accuracy shifts significantly. Model-averaged HARMONY values for the original and pruned benchmarks are reported in parentheses in the legends.

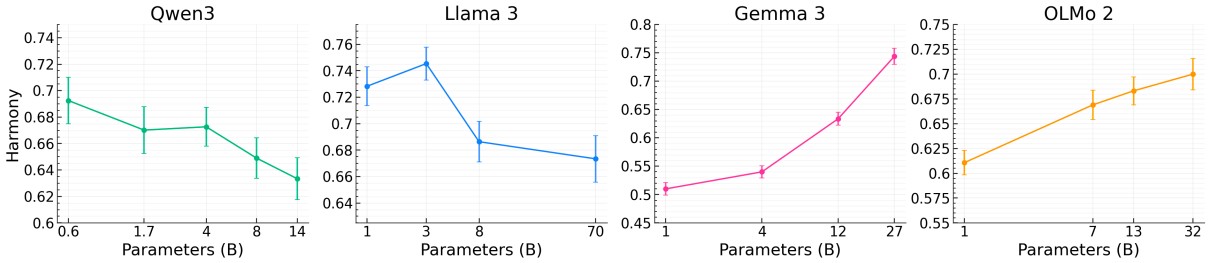


Figure 5: **Model size vs. HARMONY.** Scaling trends are *family-specific*: Qwen and Llama show negative correlations, while Gemma and OLMo show positive correlations (larger models perform more uniformly). Thus, parameter count alone is not predictive of performance uniformity. Y-axis shows each model's average HARMONY over all benchmarks.

performance uniformity.

Pre-training tokens. Under a fixed architecture, increasing the pre-training token budget leads to a rise in HARMONY after an initial dip. We examine this by tracking OLMo2 1B and OLMo2 7B across increasing token budgets. As shown in Fig. 6, HARMONY dips early and then rises steadily, while aggregate accuracy changes only slightly across checkpoints. Thus, the distribution of performance becomes more uniform even when the aggregate score remains nearly unchanged. Additional pre-training is therefore associated with a more even spread of performance across subdomains.

We emphasize that the trends reported here are empirical rather than causal, and leave the mechanisms underlying them to future work.

5 Related Work

Distributional Robustness. A large literature studies robustness to changes in the data distribution, spanning distributionally robust optimization, worst-group and group-shift robustness, invariant learning, and domain generalization (Duchi and Namkoong, 2020; Sagawa et al., 2020; Arjovsky et al., 2020; Koh et al., 2021). Across these lines of work, a common concern is that strong average performance can mask fragile behavior on minority groups, under reweightings of the training distribution, or after shifts from train to test environments (Duchi and Namkoong, 2020; Sagawa et al., 2020; Recht et al., 2019). Our work is adjacent in spirit to this literature, but differs in its object of analysis. Prior work typically asks whether a model remains

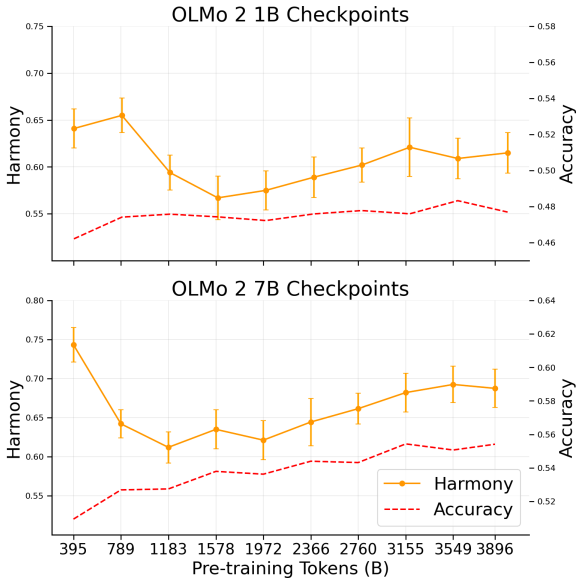


Figure 6: **Pre-training tokens vs. HARMONY.** For OLMo 2 1B/7B, HARMONY dips then steadily rises with more pre-training tokens while aggregate accuracy improves slightly.

reliable under external distribution shift (Sagawa et al., 2020; Arjovsky et al., 2020; Koh et al., 2021). By contrast, we ask whether a benchmark's aggregate score is representative of performance across its internal subdomains. In this sense, our proposal is a diagnostic for the validity of aggregate benchmark summaries. Hence, our perspective is aligned with evidence that benchmark conclusions can be sensitive to modest changes in evaluation data and test-set construction (Recht et al., 2019; Miller et al., 2021).

Assessing Benchmark Robustness. Beyond proposing new tasks, a growing body of work examines the robustness and reliability of benchmarks themselves. One line of research studies how evaluation sets behave under dynamic or adversarial conditions, including dynamic benchmarks (Kiela et al., 2021; Chiang et al., 2024) and adversarially constructed or perturbed versions (Nie et al., 2020; Croce et al., 2021). Closely related are concerns about contamination and overfitting to public benchmarks (Deng et al., 2024; Golchin and Surdeanu, 2024; Roberts et al., 2023; Dong et al., 2024), as well as redundancy, annotation artifacts, and label errors that can distort benchmark conclusions (Northcutt et al., 2021; Gema et al., 2025; Polo et al., 2024; Perlitz et al., 2024a). A complementary line of work develops frameworks and documentation practices for systematic benchmark assessment (Reuel et al., 2024; Mazumder et al., 2023; Gebru et al., 2021), while work on measurement validity asks whether benchmark scores support the inferential claims drawn from them (Yin et al., 2019; Chouldechova et al., 2024; Salaudeen et al., 2025). Other studies investigate whether benchmark gains are trustworthy, what reported scores actually capture, and how reliably comparative rankings can be stabilized (Ott et al., 2022; Dehghani et al., 2021; Singh et al., 2025; Madaan et al., 2024; Wang et al., 2024a; Heineman et al., 2025). Our work contributes to this literature by foregrounding a benchmark-level distributional perspective. Instead of treating a benchmark score as a single scalar summary, we ask whether that summary is actually representative of performance across the benchmark's internal subdomains.

Distributional Approaches to Efficient Evaluation. A separate line of work uses distributional structure in benchmarks to reduce evaluation cost while preserving comparative validity. Examples include adaptive reliability metrics (Perlitz et al., 2024b), Item Response Theory based methods for modeling item informativeness (Rodriguez et al., 2021; Tatsuoka et al., 1971), constructing compact but faithful benchmark subsets (Polo et al., 2024), and usable-information based analyses of evaluative signal (Ethayarajh et al., 2022). Although these methods differ in technique, they share the goal of exploiting benchmark structure to obtain cheaper evaluations that preserve rankings or other comparative conclusions. In contrast, we do not aim to compress evaluation, but to test whether a benchmark's aggregate score is itself a representative summary of broad competence across subdomains.

6 Conclusion

We introduce HARMONY, an entropy-based measure of how uniformly performance is distributed across a benchmark's subdomains, and use it to study *benchmark-level distributional robustness*. Mapping 19 MCQA benchmarks across five model families in the HARMONY mean-variance plane reveals substantial variation in how reliably aggregate scores reflect broad competence across subdomains. Benchmarks with high mean HARMONY and low cross-model variance yield more representative and stable evaluations, whereas low mean or high variance signal that aggregate conclusions may be fragile or model-dependent.

Our controlled analysis shows that higher-HARMONY benchmark variants tend to yield more

stable aggregate evaluation. We further find that scaling trends in HARMONY are family-specific and parameter count alone does not reliably predict performance uniformity, although larger pretraining budgets are associated with more uniform performance on average. Overall, our results show that aggregate benchmark scores can misrepresent broad competence when performance is uneven across subdomains. We therefore recommend reporting HARMONY alongside accuracy to support claims about broad model competence. We view HARMONY as complementary to, rather than a substitute for, subdomain-level analysis, as subdomain-level results reveal where strengths and weaknesses lie, while HARMONY summarizes whether aggregate benchmark performance is broadly representative across them.

Ethical Considerations

This work evaluates publicly available benchmarks and introduces *RedundantQA*, a synthetic dataset generated from author-written seeds and LLM generations with no human subjects, personal data, or sensitive attributes collected. All examples were screened to avoid offensive content and verbatim copyrighted material. We respect the licenses of all benchmarks and models used. We will release our code and RedundantQA with documentation of construction, intended use, and limitations.

HARMONY is intended to complement (and *not replace*) standard accuracy and robustness analyses. Potential risks include misinterpretation of HARMONY or benchmark rebalancing to mask undesired failure modes. We therefore report both HARMONY and accuracy and encourage transparent, multi-metric evaluation. Our experiments rely on inference with open-source models. The authors declare no conflicts of interest or external sponsorship that could bias this work.

Limitations

Our study has several limitations. First, we focus on multiple-choice benchmarks, where accuracy and predictive distributions are easy to compute, although we additionally validate HARMONY on two non-MCQA tasks. Free-form generation requires reliable scoring and may introduce judge-related uncertainty. Second, HARMONY depends on the induced subdomain partition. Although we validate predictive similarity and include model-agnostic sanity checks, the clusters may still be sensitive to the similarity metric, clustering procedure, and choice of k. Third, HARMONY is a scalar diagnostic rather than a complete measure of benchmark quality: high HARMONY does not rule out contamination, annotation artifacts, narrow coverage, or real-world mismatch, and low HARMONY does not make a benchmark unusable. Finally, our empirical findings are conditioned on the benchmarks and open model families studied, and the pruning experiments should be viewed as diagnostic interventions rather than prescriptions for benchmark construction.

Acknowledgments

AU is supported by JHU PURA (the Provost's Undergraduate Research Award) and Pistritto Research Fellowship. DK is in part supported by Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001125C0304, ONR grant (N0001424-1-2089) and JHU Provost Discovery Award (2025–2027). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA. We acknowledge the use of computational resources on the Johns Hopkins Data Science and AI Institute (DSAI) cluster.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shun-ichi Amari. 2016. *Information Geometry and Its Applications*, 1st edition. Springer Publishing Company, Incorporated.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *Preprint*, arXiv:1905.13319.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization. *Preprint*, arXiv:1907.02893.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2025. Arc prize 2024: Technical report. *Preprint*, arXiv:2412.04604.

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A. Feder Cooper, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. 2024. A shared standard for valid measurement of generative ai systems' capabilities, risks, and impacts. *Preprint*, arXiv:2412.01934.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. Robustbench: a standardized adversarial robustness benchmark. *Preprint*, arXiv:2010.09670.

Imre Csiszár and Paul C. Shields. 2004. Information theory and statistics: a tutorial. *Commun. Inf. Theory*, 1(4):417–528.

Google DeepMind. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *Preprint*, arXiv:2107.07002.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. *Preprint*, arXiv:2311.09783.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

John Duchi and Hongseok Namkoong. 2020. Learning models with uniform performance via distributionally robust optimization. *Preprint*, arXiv:1810.08750.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. *Preprint*, arXiv:2110.08420.

Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Preprint*, arXiv:2101.02235.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. *Preprint*, arXiv:2308.08493.

Aaron Grattafiori and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M. Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel J. Kochenderfer. 2024. More than marketing? on the information value of ai benchmarks for practitioners. *Preprint*, arXiv:2412.05520.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. 2025. Signal and noise: A framework for reducing uncertainty in language model evaluation. *Preprint*, arXiv:2508.13144.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. 2025. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *Preprint*, arXiv:2502.06453.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? *Preprint*, arXiv:2310.06770.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. 2025. Prismatic synthesis: Gradient-based data diversification boosts generalization in llm reasoning. *Preprint*, arXiv:2505.20161.

Aisha Khatun and Daniel G. Brown. 2024. Trutheval: A dataset to evaluate llm truthfulness and reliability. *Preprint*, arXiv:2406.01855.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in nlp. *Preprint*, arXiv:2104.14337.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, and 4 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. *Preprint*, arXiv:2012.07421.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *Preprint*, arXiv:2406.10229.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, and 26 others. 2023. Dataperf: Benchmarks for data-centric ai development. *Preprint*, arXiv:2207.10062.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018a. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018b. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. *Preprint*, arXiv:2107.04649.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *Preprint*, arXiv:1910.14599.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *Preprint*, arXiv:2103.14749.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024a. Efficient benchmarking of language models. *Preprint*, arXiv:2308.11696.

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

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1090 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? *Preprint*, arXiv:1902.10811.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Preprint*, arXiv:2411.12990.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. Data contamination through the lens of time. *Preprint*, arXiv:2310.10628.

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

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. *Preprint*, arXiv:2405.10938.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Preprint*, arXiv:1911.08731.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *Preprint*, arXiv:2310.18018.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to meaning: A validity-centered framework for ai evaluation. *Preprint*, arXiv:2505.10573.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019a. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *Preprint*, arXiv:1904.09728.

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

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

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *Preprint*, arXiv:2009.10795.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. *Preprint*, arXiv:1909.03553.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Maurice M. Tatsuoka, Frederic M. Lord, Melvin R. Novick, and Allan Birnbaum. 1971. Statistical theories of mental test scores. *Journal of the American Statistical Association*, 66:651.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. *Preprint*, arXiv:2309.08638.

Sida I. Wang, Alex Gu, Lovish Madaan, Dieuwke Hupkes, Jiawei Liu, Yuxiang Wei, Naman Jain, Yuhang Lai, Sten Sootla, Ofir Press, Baptiste Rozière, and Gabriel Synnaeve. 2024a. Eval-Arena: noise and errors on llm evaluations. `https://github.com/crux-eval/eval-arena`.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *Preprint*, arXiv:1707.06209.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao. 2025. The second half. https://ysymyth.github.io/The-Second-Half/. Written on April 10, 2025. Accessed: 2026-03-26.

Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Zhiyuan Zeng, Yizhong Wang, Hannaneh Hajishirzi, and Pang Wei Koh. 2025. Evaltree: Profiling language model weaknesses via hierarchical capability trees. *Preprint*, arXiv:2503.08893.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

A RedundantQA

To rigorously evaluate the discriminative power of similarity metrics, we construct RedundantQA, a controlled benchmark designed to disentangle genuine semantic similarity from superficial lexical overlap. Each set in RedundantQA consists of a reference question accompanied by two *true-similar* and two *false-similar* questions. The true-similar questions are paraphrases that evaluate the same underlying knowledge as the reference, while differing in surface form.[7] In contrast, the false-similar questions exhibit high lexical similarity to the reference but target distinct conceptual content. This design ensures that strong similarity metrics must go beyond surface-level cues, rewarding semantic alignment while ignoring spurious correlations.

In this section, we detail the construction (A.1) and validation (A.2) of RedundantQA, and showcase examples (A.3).

[7]E.g., variations in vocabulary, syntax, or phrasing.

A.1 Construction

We construct RedundantQA through a two-phase pipeline followed by strict validation: **(i) Seed Set Selection.** We begin by manually authoring three high-quality reference questions across four domains (Biology, Economics, Popular Culture, History). For each reference question, we also craft two paraphrases that target the same underlying knowledge (*true-similar*) and two distractors that share surface tokens but probe different concepts (*false-similar*). **(ii) Generative Expansion.** Using the seed sets as in-context learning examples, we prompt `Gemini-2.0-flash` (DeepMind, 2023) to generate 100 sets that consist of one reference question, two true-similar questions, and two false-similar questions for each domain. For different domains, we use a fixed template (Listing A.1) with domain-specific examples. This pipeline yields a large, automatically generated candidate pool.

Prompt for Generating RedundantQA (Biology)

Come up with question sets. Each set must contain:

- A reference question,
- Two same-meaning questions: These should require the same factual answer and test the same biological concept as the reference question, but they should use different wording, phrasing styles, and sentence structures.
- Two distractor questions: These should look superficially very similar to the reference question but evaluate a different knowledge or skill with different answers than the reference question.

Notes:

- Focus on the domain of biological knowledge.
- Same-meaning questions should preserve deep semantic equivalence but vary stylistically. These must have the same answer.
- Distractor questions should maximize shallow textual similarity (e.g., shared nouns, verbs, syntactic patterns) while changing the underlying meaning. So, distractor questions should trick an incapable similarity measure into thinking they are similar.

Examples:
Set 1

- *Reference Question:* What organ pumps blood throughout the human body?
- *Same-meaning Question 1:* Which organ circulates blood to deliver oxygen and nutrients?
- *Same-meaning Question 2:* What body system structure maintains blood flow across the body?
- *Distracting Question 1:* What organ removes carbon dioxide from the blood?
- *Distracting Question 2:* What organ transports nutrients through the blood?

Set 2

- *Reference Question:* What process converts glucose into energy in cells?
- *Same-meaning Question 1:* Which process produces ATP from sugar molecules?

- *Same-meaning Question 2:* What pathway transforms glucose into usable cellular energy?
- *Distracting Question 1:* What process stores glucose in cells?
- *Distracting Question 2:* What process breaks down proteins for energy?

Set 3

- *Reference Question:* What type of blood vessel carries blood away from the heart?
- *Same-meaning Question 1:* Which vessels transport oxygenated blood from the heart?
- *Same-meaning Question 2:* What structures move blood outward from the heart?
- *Distracting Question 1:* What type of blood vessel brings blood to the heart?
- *Distracting Question 2:* What blood vessel type filters blood in the kidneys?

A.2 Validation

We validate each set generated by `Gemini-2.0-flash` through a two-stage pipeline: (a) an automated and simple consistency check using `Gemini-2.0-flash` to confirm that true-similar paraphrases produce identical answers while false-similar distractors yield divergent ones (using Listing A.2); and (b) a manual review by expert annotators to correct any misclassifications, formatting issues, or errors introduced during automated filtering. After the validation step, we obtain 71, 39, 72, and 73 sets from Biology, Economics, Culture, and History domains respectively, with each set consisting of one reference question, two true-similar questions, and two false-similar questions.

This procedure yields a benchmark in which effective similarity metrics must discriminate semantic equivalence from mere lexical coincidence.

Prompt for Validating RedundantQA

Do the following questions have the same answer?
Output only yes or no.
Question 1: REFERENCE_QUESTION
Question 2: TRUE_SIM_1

A.3 Examples

We provide examples from RedundantQA across all four domains in Table 1.

B Predictive Similarity

B.1 Alternative Baselines

To better situate our predictive similarity metric relative to existing work, we first position it against a diverse set of alternative similarity measures from the literature. These baselines span surface-form (n-gram), embedding-based, gradient-based, model-output-based similarities that are commonly used in prior work. In contrast, to the best of our knowledge, predictive similarity is the only metric that explicitly leverages the full probability distributions produced by the models under evaluation.

In this section, we describe the alternative baselines compared against predictive similarity across a range of controlled settings.

Bigram. We compute an n-gram-overlap Jaccard similarity matrix. For each text x_i, we lowercase and split on whitespace, then form the set G_i of contiguous bigrams. The pairwise similarity is $S_{ij} = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}$ with $S_{ii} = 1$ for all i.[8] The resulting $S \in [0,1]^{N \times N}$ is symmetric and measures surface-form overlap.

BERTScore F1. We use BERTScore (Zhang et al., 2020) to measure semantic similarity between pairs of texts by comparing their contextualized token embeddings. Tokens are greedily matched via cosine similarity to compute precision and recall, and the final sentence-sentence score is the F1 aggregate, where $F_1 = \frac{2PR}{P+R}$. We treat this F1 value as the pairwise similarity, which yields a symmetric matrix $S \in [-1,1]^{N \times N}$. We employ $\texttt{Roberta}_{\texttt{Large}}$ (Liu et al., 2019) for obtaining the contextualized token embeddings.

Input Embeddings Cosine Similarity. We map each input to a single vector and measure pairwise similarity via cosine in embedding space. We use two variants: (i) for each example, we take the last-token hidden state from the model under evaluation, ℓ_2-normalize it, and set $S_{ij} = \hat{h}_i^\top \hat{h}_j$, (ii) we encode each input with a frozen sentence-embedding model, normalize the embeddings, and compute the same cosine-based matrix. In both cases, we obtain a symmetric matrix $S \in [-1,1]^{N \times N}$. For the sentence-embedding variant, we use `MiniLM`[9] (Wang et al., 2020) and `gte-Qwen2-7B-instruct`[10] (Li et al., 2023); for the model-under-evaluation variant, we use

[8] Note that if a text has fewer than n tokens, its n-gram set is empty. In such a case, the pairwise similarity is set to be 1. However, this is not observed in practice.

[9] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[10] https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

Biology	Economics	History	Popular Culture
Reference What process converts glucose into energy in cells? A: Cellular respiration B: Photosynthesis C: Osmosis D: Transcription	**Reference** How does increased government spending affect aggregate demand? A: Increases it. B: Decreases it. C: Has no effect. D: Only affects aggregate supply.	**Reference** Who was the first president of the United States? A: George Washington B: Abraham Lincoln C: Thomas Jefferson D: John Adams	**Reference** Who played Iron Man in the Marvel Cinematic Universe? A: Robert Downey Jr. B: Chris Evans C: Hugh Jackman D: Tobey Maguire
True Similar Which process produces ATP from sugar molecules? A: Cellular respiration B: Photosynthesis C: Osmosis D: Transcription	**True Similar** What happens to total demand in economy when the government increase its spending? A: Increases it. B: Decreases it. C: Has no effect. D: Only affects aggregate supply.	**True Similar** Who assumed leadership as America's first head of state? A: George Washington B: Abraham Lincoln C: Thomas Jefferson D: John Adams	**True Similar** Which actor portrayed Tony Stark in the MCU? A: Robert Downey Jr. B: Chris Evans C: Hugh Jackman D: Tobey Maguire
False Similar What process stores glucose in cells? A: Glycogenolysis B: Gluconeogenesis C: Glycogenesis D: Glycolysis	**False Similar** How does increased government spending affect government debt? A: Increases it. B: Decreases it. C: Has no effect. D: Only affects short-term debt.	**False Similar** Who was the first vice president of the United States? A: John Adams B: Thomas Jefferson C: Alexander Hamilton D: James Madison	**False Similar** Who played Captain America in the Marvel Cinematic Universe? A: Chris Evans B: Chris Pratt C: John Krasinski D: Matt Damon

Table 1: Example sets across all domains in RedundantQA.

`microsoft/Phi-3-mini-4k-instruct`[11] (Abdin et al., 2024), which yields the best performance in Appendix B.2.

Input and Output Embeddings Cosine Similarity. We represent each input-output pair (e.g., question+answer) as a single vector by taking the last-token hidden state of the concatenated sequence from the model under evaluation. We ℓ_2-normalize these vectors and define pairwise similarity via cosine with $S_{ij} = \hat{h}_i^\top \hat{h}_j$. The resulting $S \in [-1, 1]^{N \times N}$ is symmetric and reflects similarity over both the question and its associated answer. We use `microsoft/Phi-3-mini-4k-instruct` (Abdin et al., 2024) for obtaining the hidden states, as it yields the best performance in Appendix B.2.

G-Vendi. Following (Jung et al., 2025), we quantify the diversity of per–example gradients via a sketch-based spectral entropy. For each example, we form a compact count-sketch of the gradient of the (negative) log-probability of the correct answer under a proxy LM, yielding a matrix $G \in \mathbb{R}^{N \times d}$. We compute $C = \frac{1}{N} G^\top G$ and its eigenvalues $\{\lambda_i\}$. Let $p_i = \lambda_i / \sum_j \lambda_j$; the *G-Vendi* score is the exponential Shannon entropy of this spectrum:

$$\text{G-Vendi} \;=\; \exp\Big(-\sum_i p_i \log p_i\Big),$$

which acts as an effective rank where higher values indicate gradients spread across more orthogonal directions and lower values indicate concentration in a low dimensional subspace. For pairwise similarity, we ℓ_1-normalize the sketch rows of G and take their dot products to obtain a symmetric similarity matrix with unit diagonal. Following the original implementation, we employ `Qwen2.5-0.5B-Instruct` (Qwen et al., 2025) as the proxy model.

[11] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Method	Duplicate Catch Ratio (↑)
N-gram & Token	
Bigram	96.3
BERTScore F1	100.0
Embedding-Based	
Input Embeddings$_{\text{MiniLM}}$	100.0
Input Embeddings$_{\text{gte-Qwen2}}$	100.0
Input Embeddings$_{\text{phi3}}$	100.0
Input+Output Embeddings$_{\text{phi3}}$	100.0
Literature	
G-Vendi	100.0
CORRS$_{\text{Llama}}$	100.0
CORRS$_{\text{all}}$	100.0
IRT Representation	100.0
Predictive Similarity	100.0

Table 2: **Validation of our metric implementations.** All metrics other than bigram similarity perfectly detect exact duplicate questions, satisfying the minimum requirement to ensure implementation accuracy.

CORRS. Following (Vivek et al., 2024), given a bank of source models, we map each input i to a vector $v_i \in \mathbb{R}^M$ whose m-th entry is the logit of the probability that model m assigns to the correct choice. Then, the similarity between two examples is defined as the Pearson correlation of these vectors with $S_{ij} = \text{corr}(v_i, v_j)$. The resulting $S \in [-1, 1]^{N \times N}$ is symmetric and represents the cross-model agreement in correct class confidence across inputs. We instantiate the source bank using

the Llama model family (Grattafiori and et al, 2024) and the full set of models used in our experiments.

IRT Representation. Following (Polo et al., 2024), from a bank of source models, we form a binary response matrix $Y \in \{0,1\}^{L \times N}$ whose (ℓ, i) entry indicates whether model ℓ answered example i correctly. We then fit a d-dimensional IRT model with per-example parameters $(\alpha_i \in \mathbb{R}^d,\ \beta_i \in \mathbb{R})$ and per-model ability vectors $\theta_\ell \in \mathbb{R}^d$, using

$$\Pr(Y_{\ell i} = 1) = \sigma\big(- \theta_\ell^\top \alpha_i + \beta_i \big).$$

Here, optimization alternates between gradient updates for θ (with ℓ_2 regularization and recentring) and logistic regressions to update (α_i, β_i). Finally, we obtain the embedding $E_i = \big[\alpha_i; \beta_i\big] \in \mathbb{R}^{d+1}$ and define pairwise similarity by cosine similarity as $S_{ij} = \frac{E_i^\top E_j}{\|E_i\|\,\|E_j\|}$, which yields a symmetric matrix $S \in [-1,1]^{N \times N}$. We use $d = 200$ and instantiate the source bank using the full set of models used in our experiments.

As open-source implementations of efficient evaluation metrics are unavailable, we re-implement them following the specifications in prior work. We then validate our implementations via a sanity check in which each metric was tasked with detecting verbatim duplicate questions. As shown in Table 2, all metrics (with the exception of bigram similarity) achieve perfect performance, satisfying the minimum requirement to ensure implementation accuracy.

B.2 Measuring Semantic Similarity

An effective similarity measure for uncovering underlying data distributions must exhibit strong discriminative power, reliably identifying semantically similar data points while rejecting distractors. We empirically validate that predictive similarity meets this criterion, as it consistently distinguishes true semantic matches from misleading surface-level overlaps in RedundantQA.

As shown in Table 3, predictive similarity achieves the highest retrieval of true semantic matches across all domains, while maintaining one of the lowest rates of false matches. This indicates that it captures semantic equivalence without being misled by superficial lexical similarity. In contrast, embedding-based and bigram baselines suffer from high false positives, conflating surface-level resemblance with meaning. Metrics from the efficient evaluation literature show stronger performance but still fall short of predictive similarity. Overall, these results highlight the unique discriminative advantage of predictive similarity in measuring semantic similarity.

B.3 Inducing the Semantic Partition

A core requirement for our work is that the similarity function should induce a semantic partition of the data. We evaluate this property by inducing cluster assignments from each metric and measuring agreement with the ground-truth domain labels in RedundantQA and its reference-only subset (RedundantQA-Ref) using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). In addition, we apply the same protocol to MMLU high school subtasks, testing whether clusters recover canonical subject domains (e.g., computer science, biology, physics). Finally, we validate the top-performing similarity baselines from MMLU and RedundantQA on GPQA and M3Exam (Zhang et al., 2023).

As shown in Table 4 and 5, predictive similarity achieves the highest agreement on all three sets, while strong embedding-based baselines are competitive yet consistently behind. By contrast, token/ngram measures and metrics from the efficient evaluation literature fail to recover domain structure, indicating that they are unreliable for semantic grouping. Taken together with the pairwise retrieval evidence, these results show that predictive similarity not only discriminates true semantic matches from distractors, but also organizes instances into compact and consistent clusters. This behavior is precisely what enables a cluster-centric analysis of benchmarks, yielding low within-cluster variance and high between-cluster separation.

B.4 Capturing Benchmark HARMONY

We evaluate all alternative similarity baselines from Appendix B.1 in the identical setting of §2.4. For each baseline, we induce clusters from its similarity matrix, compute HARMONY $H(\mathcal{G})$, and report correlations between the ground-truth HARMONY and its counterpart computed from the partition induced by each baseline. We additionally validate the top-performing similarity baselines from MMLU and RedundantQA on GPQA and M3Exam.

As shown in Fig. 7, 8 and 9, predictive similarity achieves the strongest correlation with ground-truth entropy. Similar to prior validation experiments (App. B.2, B.3), embedding-based baselines are the next-best performers but consistently lag be-

Method	True Similar (↑)					False Similar (↓)				
	Biology	Economics	Culture	History	All	Biology	Economics	Culture	History	All
N-gram & Token										
Bigram	1.4	0.0	6.9	7.0	4.5	70.0	67.6	30.6	47.9	53.7
BERTScore F1	8.6	0.0	12.5	21.1	12.4	74.3	83.8	33.3	54.9	60.3
Embedding-Based										
Input Embeddings$_{\text{MiniLM}}$	42.9	24.3	62.5	66.2	54.1	27.1	51.4	5.6	12.7	21.1
Input Embeddings$_{\text{gte-Qwen2}}$	18.6	18.9	23.6	39.4	26.9	38.6	51.4	26.4	22.5	33.5
Input Embeddings$_{\text{phi3}}$	21.4	5.4	16.7	36.6	22.7	47.1	75.7	37.5	31.0	45.5
Input+Output Embeddings$_{\text{phi3}}$	51.4	62.2	40.3	56.3	52.9	22.9	27.0	16.7	15.5	20.2
Literature										
G-Vendi	62.9	37.8	36.1	46.5	47.9	5.7	2.7	6.9	5.6	5.8
CORRS$_{\text{Llama}}$	2.9	8.1	9.7	14.1	9.1	1.4	0.0	1.4	1.4	1.2
CORRS$_{\text{all}}$	35.7	13.5	59.7	59.2	47.5	1.4	5.4	0.0	1.4	1.7
IRT Representation	1.4	0.0	1.4	0.0	0.8	1.4	0.0	0.0	2.8	1.2
Predictive Similarity	**80.0**	**86.5**	**66.6**	**73.2**	**77.7**	1.4	2.7	0.0	4.2	2.1

Table 3: Proportion of identified true-similar (↑) and false-similar (↓) pairs by method and domain.

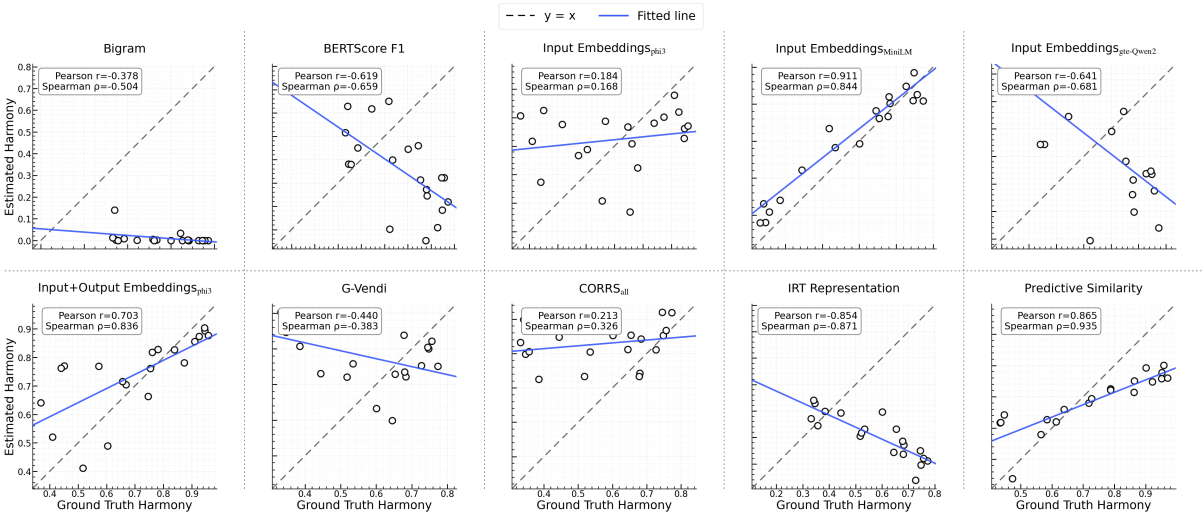


Figure 7: **Validation of our partition induction method on RedundantQA**. Predictive similarity achieves the strongest correlation between the ground truth HARMONY and estimated HARMONY, while the `MiniLM` input-embedding baseline is a close runner-up.

Method	RedundantQA		RedundantQA-Ref		MMLU-HS	
	ARI	NMI	ARI	NMI	ARI	NMI
N-gram & Token						
Bigram	-0.3	2.5	-0.4	6.7	0.1	6.2
BERTScore F1	-0.2	1.6	2.2	8.6	1.4	7.9
Embedding-Based						
Input Embeddings$_{\text{MiniLM}}$	55.8	64.4	59.4	68.4	47.4	56.8
Input Embeddings$_{\text{gte-qwen2-7b-instruct}}$	28.6	34.9	36.3	45.9	27.1	34.4
Input Embeddings$_{\text{phi3}}$	27.8	37.1	33.0	45.6	25.1	31.6
Input+Output Embeddings$_{\text{phi3}}$	57.1	67.0	58.9	70.3	51.3	59.6
Literature						
G-Vendi	6.5	11.3	4.7	10.9	3.2	9.2
CORRS$_{\text{all}}$	6.1	8.2	4.8	8.2	2.8	8.4
IRT Representation	2.1	2.7	0.4	1.9	0.4	1.5
Predictive Similarity$_{\text{Qwen3 1.7B}}$	60.4	70.4	62.5	76.5	55.4	62.1

Table 4: Validation of our partition induction method on RedundantQA and MMLU. Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) (higher is better for both) are shown for different methods on **`RedundantQA`**, **`RedundantQA-Ref`**, and **`MMLU-HS`**. Predictive similarity consistently achieves the best domain recovery.

Method	M3Exam		GPQA	
	ARI	NMI	ARI	NMI
Input Embeddings$_{\text{MiniLM}}$	60.8	65.5	58.4	64.3
Input+Output Embeddings$_{\text{Phi3}}$	61.7	68.1	59.3	66.1
Predictive Similarity$_{\text{Qwen3 1.7B}}$	**64.7**	**71.1**	**63.1**	**69.9**

Table 5: Further validation of our partition induction method on **`GPQA`** and **`M3Exam`**. Predictive similarity achieves the best domain recovery on both benchmarks.

hind, whereas token-and-n-gram overlap measures perform substantially worse. These results establish predictive similarity as the most reliable similarity metric choice for capturing the benchmark dynamics.

Lastly, Figure 10 pools all validation points from RedundantQA, MMLU, GPQA, and M3Exam, complementing the per-benchmark results in Fig-

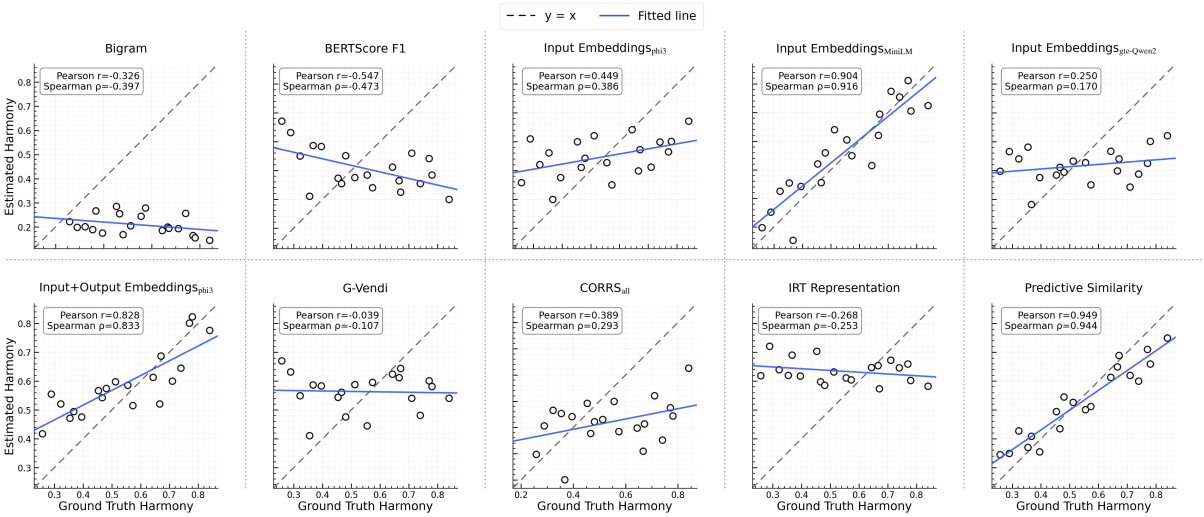
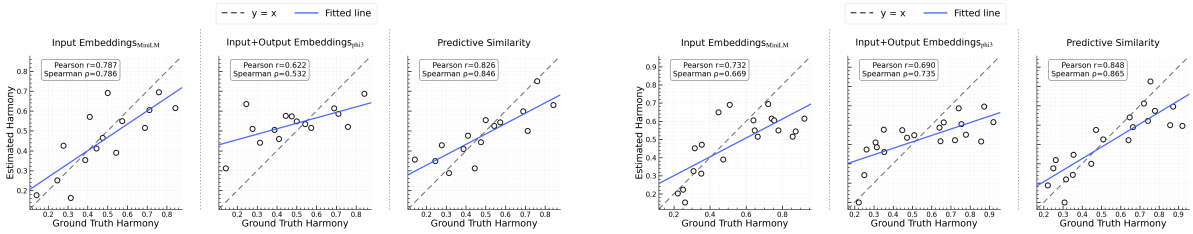


Figure 8: **Validation of our partition induction method on MMLU**. Similar to the results on RedundantQA, predictive similarity achieves the strongest correlation between the ground truth HARMONY and estimated HARMONY, while the `MiniLM` input-embedding baseline is a close runner-up.



(a) GPQA (b) M3Exam

Figure 9: **Validation of our partition induction method on GPQA and M3Exam**. Similar to the results on RedundantQA and MMLU, predictive similarity achieves the strongest correlation between the ground truth HARMONY and estimated HARMONY.

ure 7, 8, and 9, where predictive similarity consistently achieves the strongest alignment with ground-truth HARMONY across imbalance variants. When we combine these experiments into a single pool, the relationship remains strongly positive (Pearson $r = 0.869$ and Spearman $\rho = 0.879$), showing that our estimator preserves HARMONY ordering not only within each benchmark family but also across benchmarks. This supports our claim that estimated HARMONY provides a benchmark-agnostic signal of distributional robustness.

B.5 Consistency of Predictive Similarity Across Models

As defined in §2.3, predictive similarity induces, for each benchmark $\mathcal{B}$ and model f, a symmetric similarity matrix $S^{(f,\mathcal{B})} \in (0,1]^{N_\mathcal{B} \times N_\mathcal{B}}$ over the $N_\mathcal{B}$ items of $\mathcal{B}$. In this section, we ask whether these model-specific neighborhoods are idiosyncratic. To quantify cross-model consistency, we correlate the upper-triangular entries of the corresponding similarity matrices using both rank-based (Spearman) and linear (Pearson) correlations:

$$r_{\mathrm{S}}^{\mathcal{B}}(f_1, f_2) = \rho_{\mathrm{S}}\big(\mathrm{vec}(S^{(f_1,\mathcal{B})}),\ \mathrm{vec}(S^{(f_2,\mathcal{B})})\big),$$
$$r_{\mathrm{P}}^{\mathcal{B}}(f_1, f_2) = \rho_{\mathrm{P}}\big(\mathrm{vec}(S^{(f_1,\mathcal{B})}),\ \mathrm{vec}(S^{(f_2,\mathcal{B})})\big).$$

where $\mathrm{vec}(\cdot)$ stacks the upper triangle of a matrix into a vector, ρ_{S} is Spearman's rank correlation, and ρ_{P} is Pearson's correlation. We report both, as Spearman is invariant to monotone re-scalings and thus robust to calibration differences across models, while Pearson captures linear alignment in similarity magnitudes.

As shown in Table 6, predictive similarity neighborhoods exhibit substantial within-family consistency across many benchmarks. Averaging across families yields high per-benchmark means clustered in the mid-60s, with notable peaks for *LogiQA* (70.3 Spearman / 71.3 Pearson), *TruthfulQA* (72.5 / 74.5), *PIQA* (67.2 / 68.8), and *ART* (68.1 / 69.1).

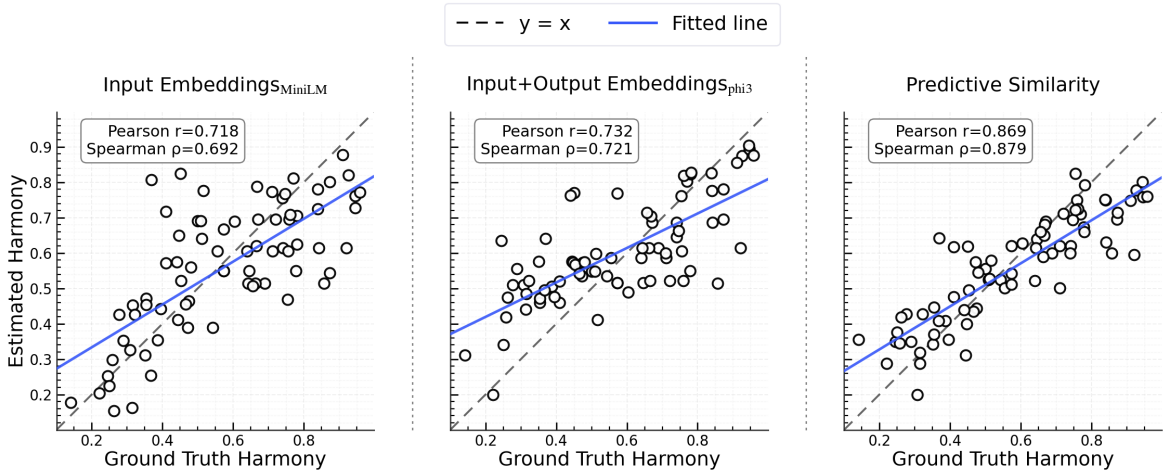


Figure 10: **Validation of our partition induction method on a combination of RedundantQA, MMLU, GPQA, and M3Exam**. Similar to the individual results, predictive similarity achieves the strongest correlation between the ground truth HARMONY and estimated HARMONY.

	Gemma		Llama		OLMo		Phi		Qwen		Average	
Benchmark	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**
AQUA-RAT	61.8	62.7	63.3	65.9	47.0	49.3	49.8	50.9	76.1	76.5	59.6	61.1
ARC-Challenge	41.9	42.7	76.3	76.9	72.9	73.9	71.9	73.9	74.5	76.1	67.5	68.7
ARC-Easy	41.6	42.0	66.6	67.1	67.9	68.7	75.3	76.3	62.7	64.0	62.8	63.6
ART	25.5	26.4	84.1	84.6	76.6	77.9	78.6	79.8	75.7	76.8	68.1	69.1
BoolQ	25.5	25.9	58.3	59.8	35.3	36.3	18.2	18.9	27.5	28.9	33.0	34.0
CommonsenseQA	33.8	35.6	33.6	34.4	62.8	64.5	42.4	43.8	39.1	39.8	42.3	43.6
COPA	27.9	29.5	76.7	78.4	68.7	70.2	64.8	67.2	51.4	52.8	57.9	59.6
GPQA	77.8	78.0	53.4	51.9	64.1	65.2	82.1	83.0	85.5	86.1	72.6	72.8
LogiQA	55.4	57.3	77.7	79.8	67.8	67.2	70.3	72.1	80.1	80.3	70.3	71.3
MathQA	33.5	34.2	61.0	62.0	55.2	56.0	54.9	54.0	57.1	56.4	52.3	52.5
OpenBookQA	33.2	34.0	73.9	75.0	70.1	72.3	74.0	75.5	71.0	72.5	64.4	65.9
PIQA	38.6	39.6	78.6	80.2	77.3	79.4	67.0	68.2	74.6	76.4	67.2	68.8
PubMedQA	50.5	50.0	60.8	60.8	37.8	36.8	50.6	50.6	63.0	61.1	52.5	51.9
QUARTZ	37.9	39.0	81.4	80.3	65.9	63.9	76.0	74.9	70.0	68.1	66.2	65.2
SciQ	22.0	22.5	55.7	58.3	60.3	61.4	56.0	56.0	58.0	60.3	50.4	51.7
SocialIQA	20.7	21.3	72.9	73.5	64.6	65.6	63.5	64.9	61.2	62.0	56.6	57.5
StrategyQA	3.6	3.7	48.2	49.6	20.5	20.9	21.6	21.4	27.9	28.9	24.4	24.9
TruthfulQA	50.4	52.7	78.4	80.7	71.4	73.6	85.3	86.4	77.1	79.3	72.5	74.5
Average	37.9	38.7	66.7	67.7	60.3	61.3	61.2	62.1	62.9	63.7	57.8	58.7

Table 6: **Cross-model consistency of predictive similarity**: Spearman and Pearson correlations (values $\times 100$) between the upper-triangle entries of affinity matrices, by benchmark (rows) and model family (columns). The rightmost block reports per-benchmark averages across families; the bottom row reports per-family averages across benchmarks; the bottom-right cell shows overall means.

Family-wise averages further show broad stability for Llama (66.7 / 67.7), Qwen (62.9 / 63.7), and Phi (61.2 / 62.1), with OLMo close behind (60.3 / 61.3) and Gemma lower (37.9 / 38.7). In particular, *ART*, *COPA*, *LogiQA*, *PIQA*, and *TruthfulQA* attain high agreement for Llama, Phi, and Qwen (typically 70–85), indicating that the induced item-item structure is largely task-driven rather than model idiosyncratic. Knowledge-centric *GPQA* is also strong for Phi and Qwen (82–86). By contrast, *StrategyQA*, *BoolQ*, and, to a lesser extent, *SocialIQA* show weaker agreement particularly for Gemma and Phi, suggesting greater family-specific effects on these benchmarks.

B.6 Probing the Determinants of Predictive Similarity

Dependence on the Tail. We test sensitivity to the probability tail by constructing a truncated variant that re-normalizes mass over the union of the `top-50` tokens, yielding $S^{(f,\mathcal{B})}_{\text{KL-top-50}}$, and contrasting it with the full $S^{(f,\mathcal{B})}_{\text{KL}}$.

JS vs. KL-based Similarity. Equation 2 defines S as an RBF of the Jeffreys divergence, producing *sharp, tunable* neighborhoods that strongly penalize coverage errors (as near-zeros drive J higher and S lower). Jensen-Shannon (JS) instead compares to the mixture $M = \frac{1}{2}(\bar{p}_f(x_i) + \bar{p}_f(x_j))$, yielding a bounded, tail-robust divergence that is

Benchmark	S_{JS}		$S_{\mathtt{KL\text{-}top\text{-}50}}$	
	Pearson	Spearman	Pearson	Spearman
AQUA-RAT	96.1	95.5	97.4	97.4
ARC-Challenge	95.7	95.2	98.5	98.5
ARC-Easy	95.4	95.0	98.7	98.7
ART	95.5	95.4	97.1	96.9
BoolQ	93.5	92.5	98.4	98.4
CommonsenseQA	97.0	96.7	99.5	99.5
COPA	94.7	95.2	95.1	94.9
GPQA	97.1	97.1	99.1	99.0
LogiQA	98.7	98.5	98.5	98.5
MathQA	99.1	99.4	96.4	96.3
OpenBookQA	97.6	97.7	97.5	97.6
PIQA	96.4	96.1	96.2	95.9
PubMedQA	96.8	96.7	99.6	99.6
QUARTZ	95.8	95.8	99.8	99.8
SciQ	93.6	93.3	98.8	98.9
SocialIQA	98.2	98.2	97.0	96.8
StrategyQA	99.4	99.5	97.0	97.0
TruthfulQA	94.6	94.2	98.1	97.7

Table 7: **Agreement between predictive similarity variants.** Per-benchmark correlations between $S_{\mathrm{KL}}^{(f,\mathcal{B})}$ and (i) $S_{\mathrm{JS}}^{(f,\mathcal{B})}$ and (ii) $S_{\mathtt{KL\text{-}top\text{-}50}}^{(f,\mathcal{B})}$ show that neighborhoods are driven by head probability mass and remain stable under truncation or mixture smoothing.

easier to compare across benchmarks. To place JS on the same similarity scale, we apply the same RBF transform from Equation 2 entrywise, obtaining $S_{\mathrm{JS}}^{(f,\mathcal{B})}$.

Following Appendix B.5, we compute Spearman and Pearson correlations between the upper-triangular entries of $S_{\mathrm{KL}}^{(f,\mathcal{B})}$ and, respectively, $S_{\mathtt{KL\text{-}top\text{-}50}}^{(f,\mathcal{B})}$ and $S_{\mathrm{JS}}^{(f,\mathcal{B})}$. We report per-benchmark scores where f is OLMo 2 7B. *High agreement* indicates that neighborhoods are driven by high probability mass and remain stable under truncation or mixture smoothing, while *low agreement* indicates sensitivity to tail mismatches or calibration asymmetries that Jeffreys magnifies but JS attenuates.

Across benchmarks, correlations between $S_{\mathrm{KL}}^{(f,\mathcal{B})}$ and (i) $S_{\mathtt{KL\text{-}top\text{-}50}}^{(f,\mathcal{B})}$ and (ii) $S_{\mathrm{JS}}^{(f,\mathcal{B})}$ variants are uniformly high (typically 95-99%) (Table 7). This indicates that the divergence between probability distributions generated by OLMo 2 7B is governed by head probability mass rather than the tail. Truncation preserves structure nearly perfectly as $S_{\mathtt{KL\text{-}top\text{-}50}}$ matches or exceeds S_{JS} on most tasks, while S_{JS} remains strongly aligned, reflecting robustness to calibration and coverage noise. Modest dips (e.g., COPA, PIQA) suggest settings where tail mismatches or asymmetries matter more, but overall the stability under truncation and mixture smoothing supports that S captures meaningful, head-driven divergence.

B.7 Theoretical and Computational Discussions

A Theoretical Perspective on Predictive Similarity. We measure similarity via the (symmetrized) D_{KL} between model predictive distributions because it aligns with how models differ operationally and geometrically. First, D_{KL} has a clear testing meaning, as it governs optimal error exponents in distinguishing two distributions. Hence, larger D_{KL} divergence implies that the model would more reliably tell the two inputs apart (by Stein/Chernoff asymptotics) (Cover and Thomas, 2006). Second, small D_{KL} guarantees closeness in total variation by Pinsker's inequality, implying high indistinguishability and hence high similarity for our purposes (Cover and Thomas, 2006). Third, D_{KL} is information monotone under coarse-graining, making the measure stable to relabeling or merging answer tokens/options that preserve semantics (Csiszár and Shields, 2004). Finally, locally D_{KL} induces the Fisher-Rao geometry on the probability simplex, so $\exp(-\tau\, \mathrm{D}_{\mathrm{KL}})$ behaves like a Gaussian kernel in the natural metric of the model's predictive space, yielding compact clusters of similar predictive behavior (Amari, 2016). We use the Jeffreys (symmetrized) form to remove directionality while retaining these properties.

Computational Overhead of Predictive Similarity. Predictive similarity is a *post hoc* computation, since we operate on the logits already cached from the benchmark evaluation. Hence, no additional model forward passes are required. Given these logits, we convert them to predictive distributions and evaluate the pairwise KL terms that define the similarity in Eq. 2.

The principal cost arises from forming pairwise interactions across N items, which is quadratic in N and linear in the label-space size D (i.e., $O(N^2 \cdot D)$ time). Memory is dominated by storing the evaluation logits ($O(N \cdot D)$) and the similarity matrix ($O(N^2)$). In practice, D corresponds to the size of the vocabulary of a given model and can be large. We therefore view the cost as $O(N^2 D)$ and the memory requirement as $O(N \cdot D)$. When N is large, standard remedies (e.g. blockwise evaluation) reduce peak memory without changing the definition of the metric. Overall, computing predictive similarity adds negligible *inference* overhead and modest *analysis* overhead relative to running the benchmarks themselves.

Compared to alternative baselines discussed in

Appendix B.1, predictive similarity is computationally frugal: it reuses cached logits and requires neither additional inference nor any backward passes. By contrast, embedding baselines, as well as BERTScore, invoke separate encoders (extra forward passes), G-Vendi relies on gradients (backward passes), and CORRS/IRT aggregate signals from a bank of models (multiple evaluations per item). While string-based methods such as Bigram are lightweight, they do not leverage model behavior. Thus, predictive similarity offers a favorable trade-off between compute and quality when benchmarks are already being run.

B.8 Discussion on Model-Specific Similarity

Our goal is to *evaluate benchmarks*, not to define a single, task-agnostic neighborhood of data points. A benchmark can be distributionally robust for one model but not for another. Accordingly, the similarity function used to induce the partition $\mathcal{G}_f$ should be *conditional on the model* f. We provide our rationale below.

Evaluation target is $H_{\mathcal{B}}(f)$. Section 2 defines HARMONY *per model*, $H_{\mathcal{B}}(f)$, and then aggregates across f via (μ_H, σ_H^2) in Eq. 1. Using a *global, model-agnostic* similarity collapses distinct predictive neighborhoods into a single partition, implicitly assuming that $\mathcal{G}_f$ is invariant across f. This undermines the very statistic we report: two models with the same accuracy profile but different predictive structure could receive the same H under a fixed partition, obscuring model specialties.

Benchmarks are instruments relative to a model. A benchmark is a diagnostic instrument for a *given* model: priors, tokenization, calibration, and pretraining exposure all change which items are *similar* from the model's perspective. Hence, a model can perform uniformly on a benchmark while another one overfits to certain subdomains. Model-specific similarity preserves this relativity, letting robustness vary meaningfully across families.

C Design Choices for HARMONY

We briefly motivate our definition of HARMONY and our choice of normalized Shannon entropy. The primary reason is pragmatic: Shannon entropy is the canonical measure of uncertainty (and thus uniformity) in information theory, so its behavior and scale are well understood and easy to interpret.

Recall that we first compute non-negative weights $\pi_i = w_i K_i$ over subdomains $i = 1, \ldots, K$ that increase with (i) the mass of the dataset assigned to subdomain i and (ii) how close the model's performance on i is to the benchmark mean. Let

$$p_i \;=\; \frac{\pi_i}{\sum_{j=1}^{K} \pi_j}, \qquad \sum_{i=1}^{K} p_i = 1,$$

denote the normalized weights as in Section 2.2, so that $p = (p_1, \ldots, p_K)$ yields a probability vector that summarizes *where* in the benchmark the model is performing well.

Our design is guided by three criteria:

(i) **Joint coverage and uniformity:** HARMONY should increase when more dataset mass lies on subdomains whose performance is close to the benchmark mean, and decrease when performance is concentrated on a few subdomains.

(ii) **Comparability across benchmarks:** HARMONY should lie in $[0, 1]$ so benchmarks with different numbers of subdomains are directly comparable.

(iii) **Interpretability as robustness:** High HARMONY should indicate that the benchmark behaves *as if* many subdomains are near the mean, while low HARMONY signals effective concentration on a smaller subset.

We therefore define HARMONY as the normalized Shannon entropy:

$$H(\mathcal{G}_f) \;=\; -\frac{1}{\log k} \sum_{i=1}^{k} p_i \log\bigl(p_i + \varepsilon\bigr) \in [0, 1],$$

This choice yields:

- **Boundedness:** $0 \leq H(\mathcal{G}_f) \leq 1$, with 1 attained only when $p_i = 1/k$ for all i, and values approaching 0 as mass concentrates on a single subdomain.
- **Effective number of well-performing subdomains:** Let $\tilde{H}(p) = -\sum_{i=1}^{k} p_i \log p_i$ denote the (unnormalized) Shannon entropy. Its exponential $\exp(\tilde{H}(p))$ is often interpreted as the *effective number* of equally weighted categories. Since $H(\mathcal{G}_f) = \tilde{H}(p)/\log k$, this can be read as a normalized effective number, i.e., the effective fraction of subdomains whose performance is close to the benchmark mean.

- **Sensitivity to non-uniformity:** Shannon entropy is strictly concave and Schur-concave, which makes any move from a more uniform to a more concentrated distribution $p = (p_1, \ldots, p_k)$ strictly decrease $H(\mathcal{G}_f)$.

C.1 Alternative Uniformity Measures

Variance and coefficient of variation. A natural idea is to measure dispersion of subdomain performances $\{\Psi(f; A_i)\}_{i=1}^k$ via a (weighted) variance

$$\mathrm{Var}\big(\Psi(f;\cdot)\big) = \sum_{i=1}^{k} w_i \Big(\Psi(f; A_i) - \bar{\Psi}\Big)^2,$$

$$\bar{\Psi} = \sum_{i=1}^{k} w_i \Psi(f; A_i).$$

However, variance is scale-dependent and not naturally normalized to $[0, 1]$, and it does not admit an *effective number of subdomains* interpretation. The coefficient of variation partly addresses scale but becomes unstable when the mean is close to 0 or 1, which is common for strong models. In contrast, HARMONY depends only on the probability distribution $p = (p_1, \ldots, p_k)$ over subdomains and is automatically normalized.

Gini coefficient and related indices. Inequality indices such as the Gini coefficient measure concentration but are typically defined directly on the performance values $\{\Psi(f; A_i)\}_{i=1}^k$, not on a distribution that already combines dataset mass and closeness-to-mean. This makes it harder to interpret them as describing how *dataset mass* is distributed over near-mean vs. outlier subdomains, and they lack the effective-number interpretation that entropy has.

Rényi entropies. Rényi entropies form a one-parameter family

$$H_\alpha(p) \;=\; \frac{1}{1-\alpha} \log\Big(\sum_{i=1}^{k} p_i^\alpha\Big), \quad \alpha > 0, \alpha \neq 1,$$

with Shannon entropy recovered as the limit $\alpha \to 1$. Different orders α emphasize different parts of the distribution ($\alpha > 1$ focuses on high-probability subdomains and $\alpha < 1$ emphasizes low-probability ones), effectively introducing a hyperparameter that controls how aggressively HARMONY responds to tails and modes. In this sense, choosing a particular Rényi order α is analogous to choosing a norm order (e.g., ℓ_α) when measuring distances. Therefore, Rényi entropy yields a whole family of alternatives, but also an extra design choice about which order to prefer. We deliberately avoid this additional degree of freedom. Using the Shannon case $\alpha = 1$ gives a single, parameter-free and widely interpretable notion of entropy that treats subdomains according to their weights p_i, without further reweighting.

C.2 Robustness of Harmony

HARMONY is designed to penalize benchmarks that leave small, poorly performing regions unaddressed. Ignoring the smoothing term ε for clarity, the underlying (unnormalized) Shannon entropy

$$\tilde{H}(p) = -\sum_{i=1}^{k} p_i \log p_i$$

can be viewed as the average information content under p. Each subdomain contributes $p_i \log(1/p_i)$, and $\log(1/p_i)$ grows as p_i becomes small. As a result, shifting even modest probability mass away from under-represented, well-performing subdomains (or toward poorly performing ones) leads to a noticeable drop in $H(\mathcal{G}_f)$. In this sense, HARMONY is particularly sensitive to *gaps* in coverage, whereas variance-based dispersion measures tend to be dominated by large deviations in high-mass components and can largely ignore small but problematic regions.

Our formulation is, like any scalar index, not completely immune to adversarial benchmark design (e.g., redefining the subdomain partition). However, for a fixed partition and mean performance, increasing $H(\mathcal{G}_f)$ requires genuinely redistributing dataset mass so that a larger share lies in subdomains whose performance is close to the mean, while simple rescalings of scores or adding redundant, near-duplicate subdomains do not help. This makes HARMONY behave as an interpretable robustness index with higher values corresponding to benchmarks where performance is spread broadly across the subdomains, rather than confined to certain subdomains.

D HARMONY with Model-Agnostic Partitions

Our main analysis (§3) uses model-sensitive predictive-similarity partitions. This choice is intended to capture the subdomain structure relevant to each model's behavior. As a sanity check, we

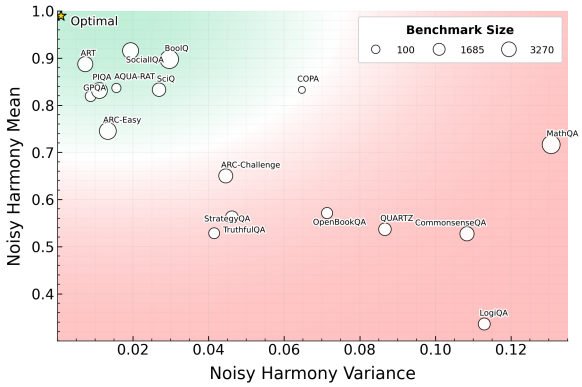


Figure 11: **MCQA benchmark landscape using model-agnostic MiniLM partitions.** Each point represents a benchmark plotted by the mean and variance of HARMONY across models, using a fixed partition induced from `all-MiniLM-L6-v2` embeddings. The qualitative structure is similar to the model-sensitive predictive-similarity analysis in Fig. 3a.

ask whether the resulting MCQA benchmark landscape changes substantially when the partition is instead fixed and model-agnostic. To this end, we repeat the MCQA analysis using partitions induced from `all-MiniLM-L6-v2` embeddings. For each benchmark, we embed all inputs with MiniLM, compute pairwise cosine similarities, and apply the same clustering procedure used in the main analysis. The resulting partition is fixed for a benchmark and reused across all models.

Figure 11 shows the resulting mean-variance plane under the MiniLM partitions. Qualitatively, the benchmark landscape remains similar to the model-sensitive analysis in Fig. 3a, as benchmarks that are comparatively high or low in HARMONY under predictive-similarity partitions tend to remain so under the model-agnostic partition. Table 8 summarizes this agreement quantitatively. Across benchmark-model pairs, HARMONY scores based on MiniLM partitions show strong agreement with the model-sensitive scores (Pearson $r = 0.897$), with a mean absolute difference of 0.109. At the benchmark level, the mean HARMONY values are also strongly correlated (Pearson $r = 0.859$), and the cross-model variances remain reasonably aligned (Pearson $r = 0.800$).

Overall, this sanity check suggests that the main MCQA conclusions are not driven solely by the model sensitivity of the partitioning step. While MiniLM partitions do not exactly reproduce the model-sensitive scores, they preserve the broad benchmark-level structure of the analysis.

Comparison level	Pearson r	MAE
Benchmark-model HARMONY	0.897	0.109
Benchmark mean μ_H	0.859	0.082
Benchmark variance σ_H^2	0.800	0.019

Table 8: Agreement between model-sensitive predictive-similarity partitions and fixed model-agnostic MiniLM partitions on MCQA benchmarks. The first row compares individual HARMONY scores over all benchmark-model pairs, while the latter rows compare benchmark-level summaries across MCQA benchmarks.

Model	IFEval		GSM8K	
	Score	HARMONY	Score	HARMONY
Qwen3-4B	76.3	0.818	79.1	0.877
Qwen3-8B	80.1	0.799	83.5	0.894
Llama-3.2-3B	69.8	0.687	75.8	0.701
Llama-3.1-8B	77.9	0.750	81.1	0.824

Table 9: HARMONY on non-MCQA benchmarks using fixed model-agnostic partitions.

E HARMONY Beyond Multiple-Choice Evaluation

We additionally evaluate HARMONY on two non-MCQA benchmarks: GSM8K for open-ended mathematical reasoning and IFEval for instruction following. We induce fixed, model-agnostic partitions using MiniLM embeddings, yielding six subdomains for GSM8K and four for IFEval. For GSM8K, we extract and normalize the final numerical answer and use exact match, while for IFEval, we use binary pass or fail judgments from Llama-3.1-70B-Instruct.

As shown in Table 9, HARMONY provides information complementary to the aggregate score across both benchmarks. For example, Qwen3-8B improves over Qwen3-4B on IFEval in aggregate score (76.3 to 80.1), while HARMONY decreases (0.818 to 0.799), indicating that the performance is less uniformly distributed across subdomains. These results demonstrate that HARMONY can be applied with task-appropriate evaluation measures beyond multiple-choice accuracy. We evaluate on the full test sets, using base Qwen checkpoints for GSM8K and instruction-tuned Qwen checkpoints for IFEval, while using instruction-tuned Llama checkpoints for both tasks. Generation follows the settings recommended in the corresponding model cards and technical reports.

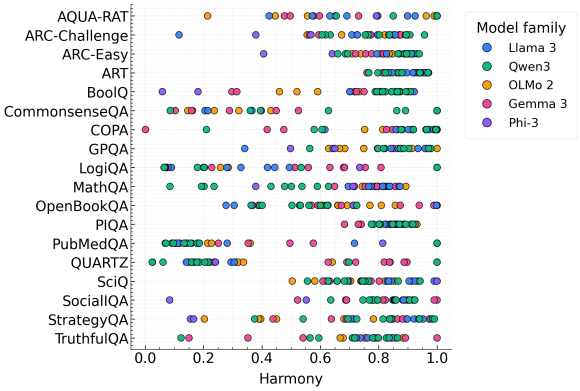


Figure 12: Model-wise decomposition of HARMONY for MCQA benchmarks.

F Model-wise Decomposition of Benchmark HARMONY

In §3.2, we position each benchmark $\mathcal{B}$ using the cross-model mean $\mu_H(\mathcal{B})$ and variance $\sigma_H^2(\mathcal{B})$. We now resolve this view at the model level. For each benchmark, Fig. 12 plots the per-model vector $\{H_{\mathcal{B}}(f)\}_{f\in\mathcal{F}}$, revealing structure that is obscured by aggregation.

Tight *horizontal* groupings (small spread across models) indicate *model-invariant* distributional balance, where different families assign similar HARMONY to the same benchmark, suggesting that aggregate accuracy reflects uniform competence irrespective of architectural or training choices. Conversely, wide horizontal scatter exposes *model-dependent robustness*, as some families concentrate performance on a few subsets (low HARMONY), while others distribute performance more evenly (high HARMONY).

We note that benchmarks with tight clusters are favorable for cross-family comparison, as accuracy rankings are less likely to be artifacts of benchmark composition. In contrast, wide scatter warns that leaderboard deltas may be driven by subsets that particular families exploit. In such cases, we suggest reporting accuracy alongside the HARMONY profiles of the models under evaluation, $\{H(\mathcal{G}_f)\}_f$.

G Assessing Benchmark Size as a Confound

A natural concern is that benchmark size might confound our mean-variance characterization of Harmony. In this section, we show analytically that HARMONY is invariant to uniform rescaling of benchmark size and empirically that benchmark size does not explain the characteristics observed in the mean-variance plane.

Recall that, given a partition $G_f = \{A_i\}_{i=1}^k$ of benchmark $\mathcal{B}$, HARMONY is computed using the relative subset weights $w_i = |A_i|/|\mathcal{B}|$ and a normalized Shannon entropy over proximity-based scores, scaled by $\log k$. Because only the normalized weights $\{w_i\}$ and the relative distribution of performance across subdomains are present in the definition, operations that uniformly scale the benchmark size (e.g., duplicating all items or uniformly sub-sampling) should leave HARMONY unchanged in principle. HARMONY is therefore designed to reflect how coverage and performance are distributed across semantic subdomains in a manner that is not impacted by the number of examples given that subdomain proportions remain the same.

To assess benchmark size as a potential confound in our setting, we first compute the Pearson correlation between benchmark size and (i) the mean of HARMONY across models and (ii) the variance of HARMONY across models for the set of 18 MCQA benchmarks in our experimental setup. We obtain:

$$\mathbb{E}_{f\sim\mathcal{F}}[H_{\mathcal{B}}(f)] \text{ vs. size: } r = -0.31, \quad p = 0.211$$

$$\mathrm{Var}_{f\sim\mathcal{F}}(H_{\mathcal{B}}(f)) \text{ vs. size: } r = 0.23, \quad p = 0.365$$

The associated p-values show no statistically significant evidence of linear dependence between benchmark size and either axis of the mean-variance plane. In particular, larger benchmarks do not exhibit systematic shifts in mean HARMONY or in cross-model variance, indicating that benchmark size does not explain the structure observed in Figure 3.

Second, we perform a controlled sub-sampling experiment on three benchmarks selected from different regions of the mean-variance plane. For each benchmark, we apply several down-sampling ratios uniformly, hence preserving the relative frequencies of semantic subdomains. We then recompute HARMONY across models. The resulting trajectories in the mean-variance plane are shown in Figure 13. Across all three benchmarks, we observe only minor fluctuations in both the mean and variance of Harmony, and the points remain within the same qualitative regions of the plane.

Taken together, the size-invariance of HARMONY under uniform rescaling, the lack of statistically significant correlations between benchmark size and HARMONY statistics, and the stability observed in the sub-sampling experiment support our view that the mean-variance plane captures

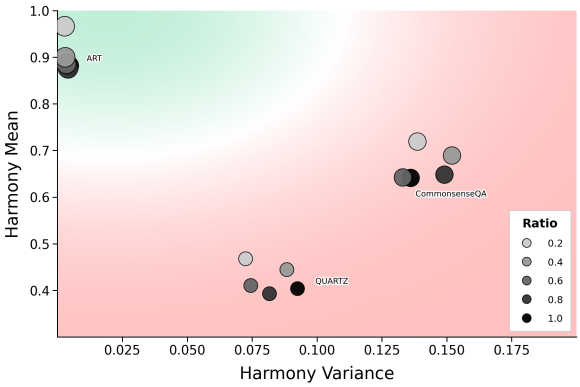


Figure 13: **Effect of benchmark size on HARMONY.** For ART, QUARTZ, and CommonsenseQA, we downsample at several ratios while preserving subdomain proportions and recompute HARMONY across models. The resulting trajectories show only minor fluctuations in both mean and variance.

differences in *distributional robustness*, rather than merely reflecting benchmark size.

H Assessing k as a Confound

Our construction of HARMONY relies on a partition of each benchmark into k semantic subdomains. A natural concern is that our choice of k might introduce a confound, if k systematically tracked HARMONY in a way that would inflate or suppress our robustness measure. In this section, we empirically study k's relationship to benchmark size and HARMONY.

As described in Sec. 2.3, for each benchmark we sweep $k \in \{2, \ldots, 20\}$ and select the value that maximizes the silhouette score, a well-established metric of clustering quality (Rousseeuw, 1987). This procedure determines k in a fully data-driven manner based on the geometry of the representation space, and k is never hand-tuned to optimize HARMONY.

To quantify how k relates to benchmark size and HARMONY, we compute Pearson correlations between (i) benchmark size and the selected k, and (ii) k and the HARMONY for each benchmark-model pair. We obtain:

$$\text{HARMONY vs. } k: \quad r = 0.274, \quad p = 0.2818,$$

$$\text{Benchmark size vs. } k: \quad r = 0.914, \quad p = 0.0000.$$

The strong correlation between benchmark size and k confirms the intuitive behavior that larger benchmarks admit more clusters. Crucially, however, there is no statistically significant linear relationship between k and HARMONY: the correlation coefficient is weak and the associated p-value indicates no evidence against the null hypothesis of zero correlation at standard significance levels.

This distinction matters because our conclusions rely on HARMONY rather than k or benchmark size itself. HARMONY already accounts for varying cluster sizes through the weights $w_i = |A_i|/|B|$, so differences in the number of clusters primarily reflect how finely the benchmark can be partitioned, not a direct change in the robustness measure. The absence of a statistically significant correlation between k and HARMONY indicates that our robustness measure is not spuriously driven by the specific value of k, even though k increases with benchmark size as expected.

I Assessing Pruning Ratio as a Confound

To explicitly rule out pruning ratio as a potential confound in the experiments of Sec. 4.1, we repeat the pruning procedure while holding the pruning budget fixed across all benchmarks. Concretely, we evaluate five ratios, $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and report the corresponding results in Fig. 14-18.

Across all pruning ratios, we observe the same pattern: (i) pruning reliably increases HARMONY, (ii) benchmarks with low HARMONY exhibit statistically significant accuracy shifts, and (iii) benchmarks with high HARMONY rarely show such shifts, even under more aggressive pruning. Although post-pruning accuracies change with the amount of data removed, these effects are stable across pruning budgets. Moreover, the magnitude of the accuracy change tends to saturate, as beyond moderate pruning (around $p = 0.3$), further increases in the pruning ratio produce only marginal additional shifts. Taken together, these results indicate that the results in Figure 4 are not explained by pruning ratio alone and are robust to the choice of pruning budget.

J Details of Statistical Significance Test

We assess whether the subset we keep after pruning has a higher mean than the full set using a nonparametric, *coupled* bootstrap sign test. Let $a_1, \ldots, a_N$ be per-example accuracy and $k_i \in \{0, 1\}$ indicate membership in the keep subset $K = \{i : k_i = 1\}$. For $b = 1, \ldots, B$ with $B = 10000$, we draw a bootstrap sample of indices $S^{(b)}$ of size N (with replacement), compute both means on the *same*

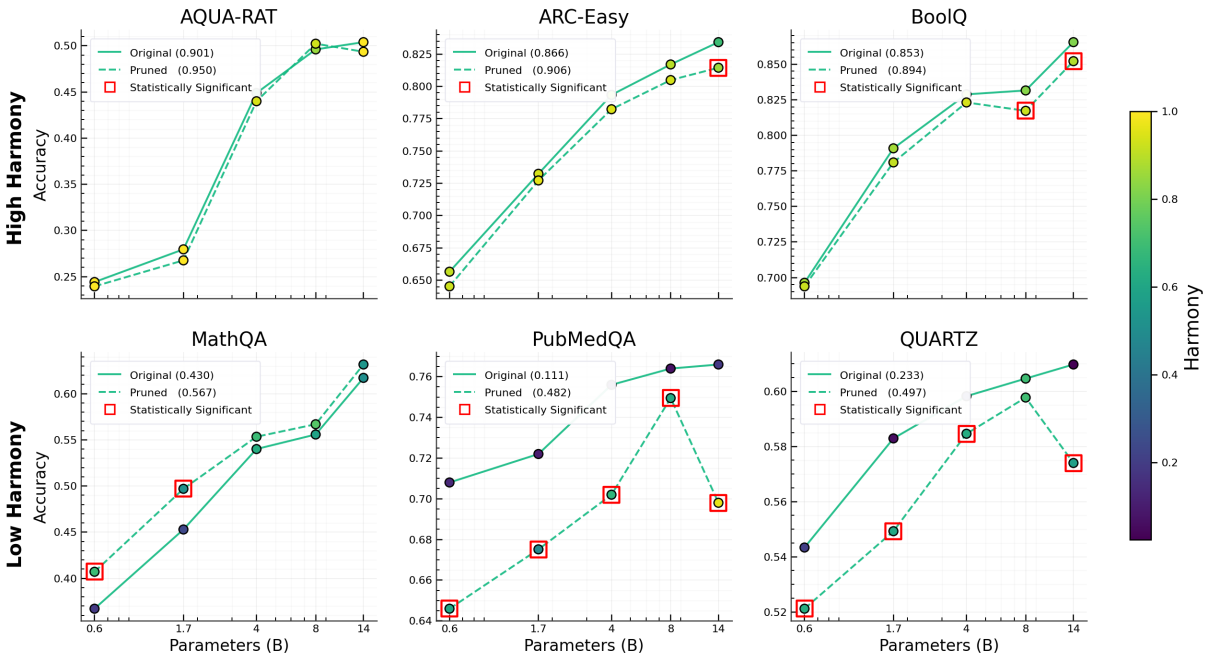


Figure 14: Pruning experiments replicated with pruning ratio $p = 0.1$ for all benchmarks.

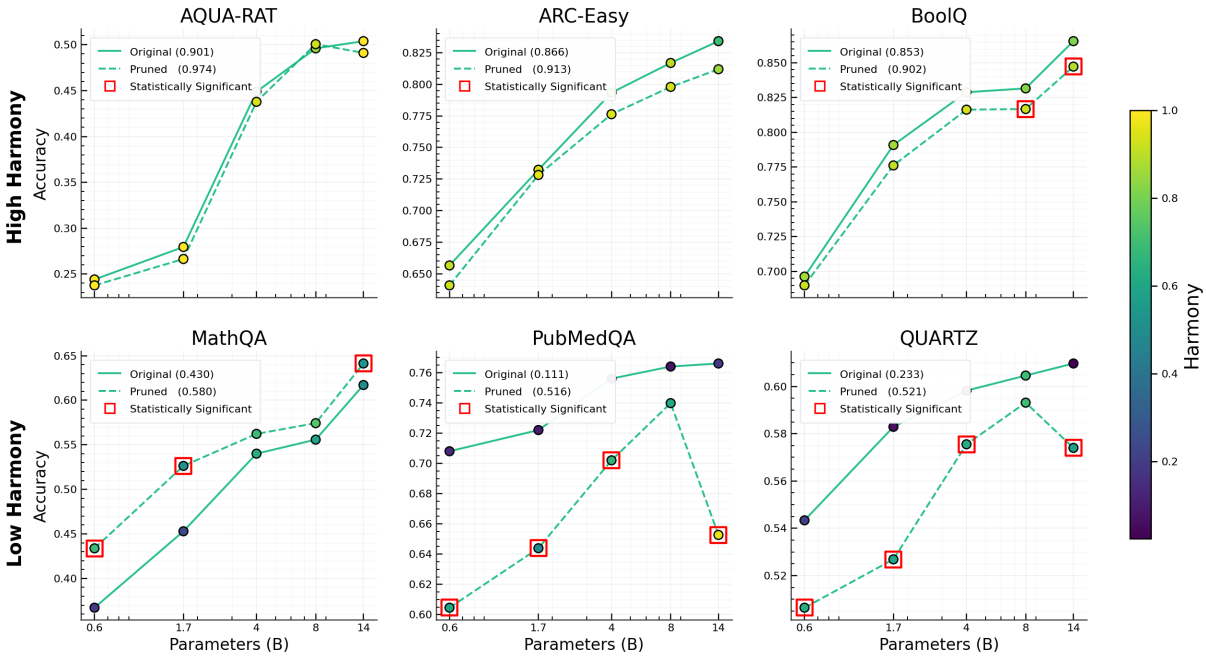


Figure 15: Pruning experiments replicated with pruning ratio $p = 0.2$ for all benchmarks.

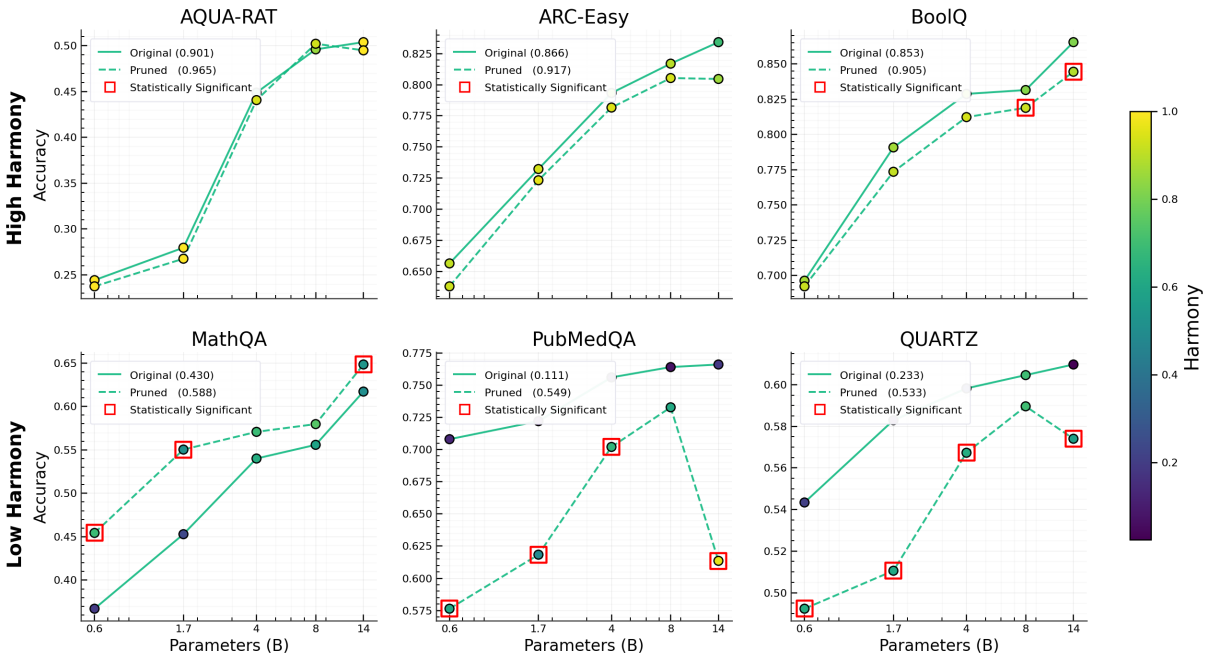


Figure 16: Pruning experiments replicated with pruning ratio $p = 0.3$ for all benchmarks.

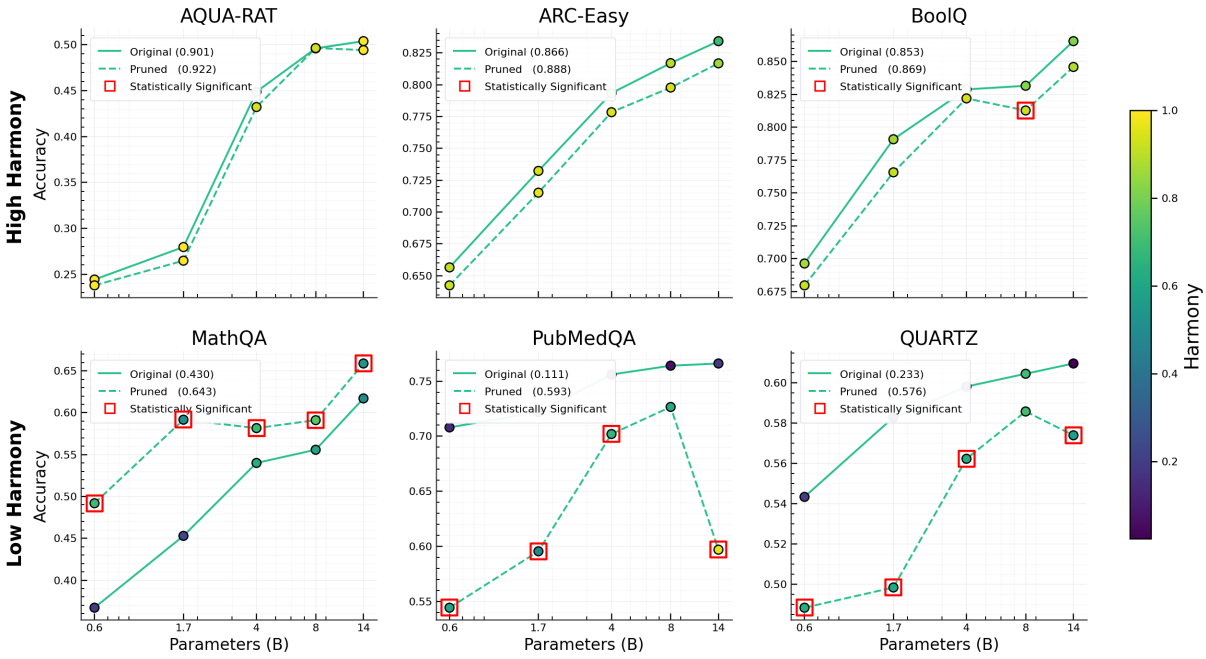


Figure 17: Pruning experiments replicated with pruning ratio $p = 0.4$ for all benchmarks.

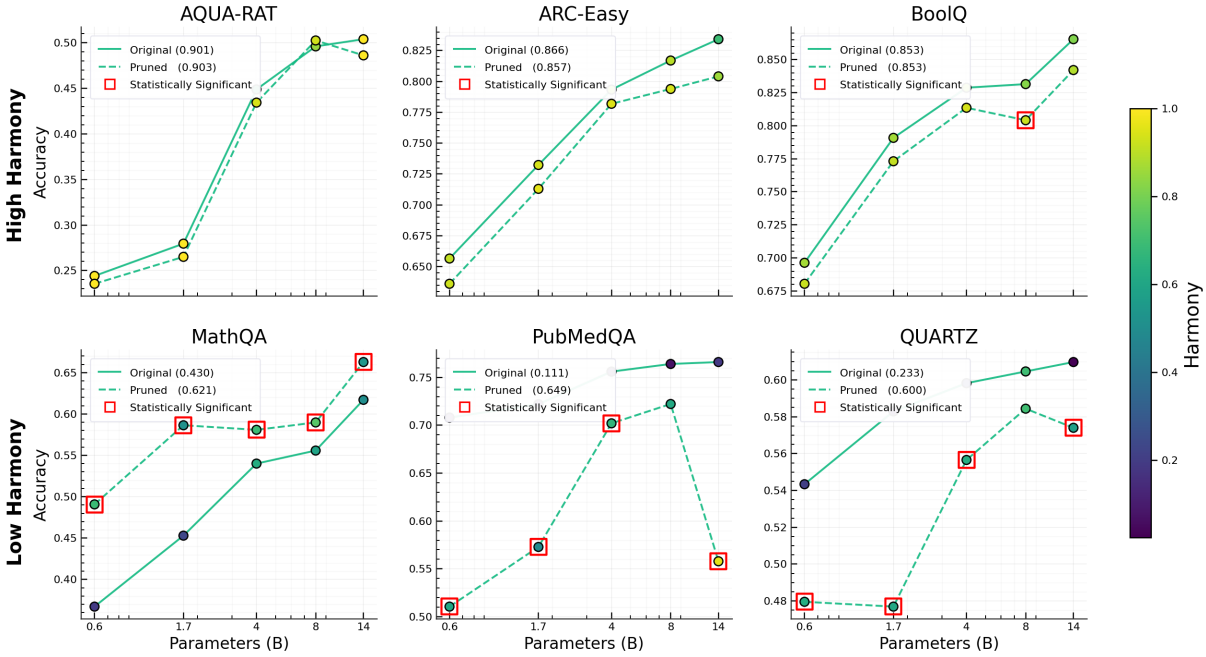


Figure 18: Pruning experiments replicated with pruning ratio $p = 0.5$ for all benchmarks.

resample, and then take their difference:

$$\bar{a}_{\text{all}}^{(b)} = \frac{1}{N} \sum_{i \in S^{(b)}} a_i, \tag{3}$$

$$n_{\text{keep}}^{(b)} = \sum_{i \in S^{(b)}} k_i, \tag{4}$$

$$\bar{a}_{\text{keep}}^{(b)} = \frac{1}{n_{\text{keep}}^{(b)}} \sum_{i \in S^{(b)}} k_i \, a_i, \tag{5}$$

$$\Delta^{(b)} = \bar{a}_{\text{keep}}^{(b)} - \bar{a}_{\text{all}}^{(b)}. \tag{6}$$

Resamples with $n_{\text{keep}}^{(b)} = 0$ are discarded (to avoid degenerate runs we cap total draws at $3B$); let $m \leq B$ be the number of valid differences retained. We then form a two-sided p-value from the sign statistic with a plus-one small-sample correction:

$$r = \sum_{b=1}^{m} \mathbf{1}\left\{\Delta^{(b)} \geq 0\right\}, \tag{7}$$

$$p = \min\left\{1,\, 2\min\left(\frac{r+1}{m+1}, \frac{m-r+1}{m+1}\right)\right\}. \tag{8}$$

We fix the random seed for reproducibility and declare significance at level $\alpha = 0.05$ when $p < \alpha$ (testing $H_0 : \mathbb{E}[\Delta] = 0$ vs. $H_1 : \mathbb{E}[\Delta] \neq 0$).

K Multi-Dimensional Evaluation

Motivated by the skewed aggregate scores in low HARMONY benchmarks, we conduct model evaluation at finer granularity. Following recent work on fine-grained evaluation (Zeng et al., 2025), we recursively induce partitions as described in §2.3. This procedure yields a *labeled tree*, where the root is the full benchmark; each internal node is a subset from the partitioning of its parent node; and leaves are atomic subdomains that admit no further valid split (see App. P for details). The resulting hierarchy enables interpretable, multi-dimensional evaluation, where each dimension corresponds to a subdomain of the benchmark.

We illustrate this approach on two examples: *MMLU College Biology* and *MMLU College Physics*. The average HARMONY across Qwen3 models is markedly higher for biology (0.8538) than for physics (0.7534). This suggests that the aggregate accuracy for biology is a more representative reflection of the performance across subdomains. Indeed, Table 10 shows that rankings in biology subdomains mirror the overall ordering: models that achieve higher overall accuracy also achieve higher accuracy in every subdomain. In other words, with one exception, models with superior overall performance are not surpassed by lower-performing models in any biology subdomain.[12] This alignment underscores that the aggregate score is a consistent and robust summary of subdomain performance in biology.

In contrast, physics exhibits lower HARMONY

[12] Comparison of 1.7B and 4B in *Molecular & Cellular Biology* is the only exception.

and more notable divergences (Table 11). For example, Qwen3-4B lags behind Qwen3-14B in overall accuracy (58.8% vs. 69.6%), yet it surpasses it in *Special Relativity* (71.4% vs. 57.1%) and *Electromagnetism* (83.3% vs. 66.7%). Similarly, Qwen3-0.6B, despite its weak overall score (24.5%), achieves competitive performance in *Quantum Mechanics* (56.5%), outperforming Qwen3-1.7B (34.8%). These cases highlight how aggregate scores can obscure areas of relative strength, and how fine-grained, multi-dimensional evaluation reveals nuanced interpretation of model competence across subdomains.

We provide extended results for Qwen3 and Gemma 3 model families across 2 MCQA benchmarks and 6 MMLU subtasks in Appendix L.

L Extended Results: Multi-Dimensional Evaluation

We provide the extended results of multi-dimensional evaluation conducted as described in Appendix K. Our setup consists of Qwen3 and Gemma 3 model families and ARC-Easy, BoolQ, MMLU Anatomy, MMLU College Biology, MMLU College Computer Science, MMLU College Mathematics, MMLU College Physics, MMLU High School US History.

L.1 Qwen3 Family

See Tables 12, 13, 14, 15, 16, 17, 18, and 19 for the extended results of the Qwen3 model family.

L.2 Gemma 3 Family

See Tables 20, 21, 22, 23, 24, 25, 26, and 27 for the extended results of Gemma 3 model family.

M Model List

We list all evaluated models and provide links to their open-source weights.

- Qwen3: Qwen3-0.6B-Base, Qwen3-1.7B-Base, Qwen3-4B-Base, Qwen3-8B-Base, Qwen3-14B-Base, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B.
- Llama 3: Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, Llama-3.1-70B, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct.
- Olmo 2: OLMo-2-0425-1B, OLMo-2-1124-7B, OLMo-2-1124-13B, OLMo-2-0325-32B, OLMo-2-0425-1B-Instruct, OLMo-2-1124-7B-Instruct, OLMo-2-1124-13B-Instruct, OLMo-2-0325-32B-Instruct.
- Gemma 3: gemma-3-1b-pt, gemma-3-4b-pt, gemma-3-12b-pt, gemma-3-27b-pt, gemma-3-1b-it, gemma-3-4b-it, gemma-3-12b-it, gemma-3-27b-it.
- Phi-3: Phi-3-mini-4k-instruct, Phi-3-medium-4k-instruct.

N Extended Results: How does model performance change with increased HARMONY?

In this section, we generalize the pruning experiments from §4.1 beyond the illustrative cases to *all* model families and benchmarks in our setup. Our aim is methodological: we examine how aggregate accuracy and per-subset dispersion evolve as we progressively rebalance a benchmark. Concretely, for each (model, benchmark) pair we sweep a pruning budget (scheduled inversely to baseline HARMONY), recompute HARMONY and accuracy at each budget, and compare the pruned-set accuracy to the full-set accuracy using the coupled bootstrap significance test detailed in App. J. Family-wise plots in this section visualize these trajectories, allowing us to observe whether increased HARMONY coincides with stable (or shifting) aggregate scores and tighter per-subset distribution of performance.

Across all model families (Fig. 19, 20, 21, 22), two patterns consistently hold. **(i) Accuracy shifts with increased HARMONY.** As pruning raises HARMONY, aggregate accuracy frequently changes in a statistically significant manner (App. J), indicating that low HARMONY composition can result in a misleading aggregate score. **(ii) Low HARMONY benchmarks are fragile.** Benchmarks starting with lower HARMONY exhibit more instances of significant accuracy change under the pruning procedure than high HARMONY benchmarks, underscoring their susceptibility to presenting misleading aggregate scores.

O Additional Related Work

Language Model Evaluation. Reliable evaluation is essential for accurately assessing model capabilities and enabling fair comparisons, which in turn informs future developments. Hence, there has been a surge in the development of benchmarks designed to test various model capabilities, such

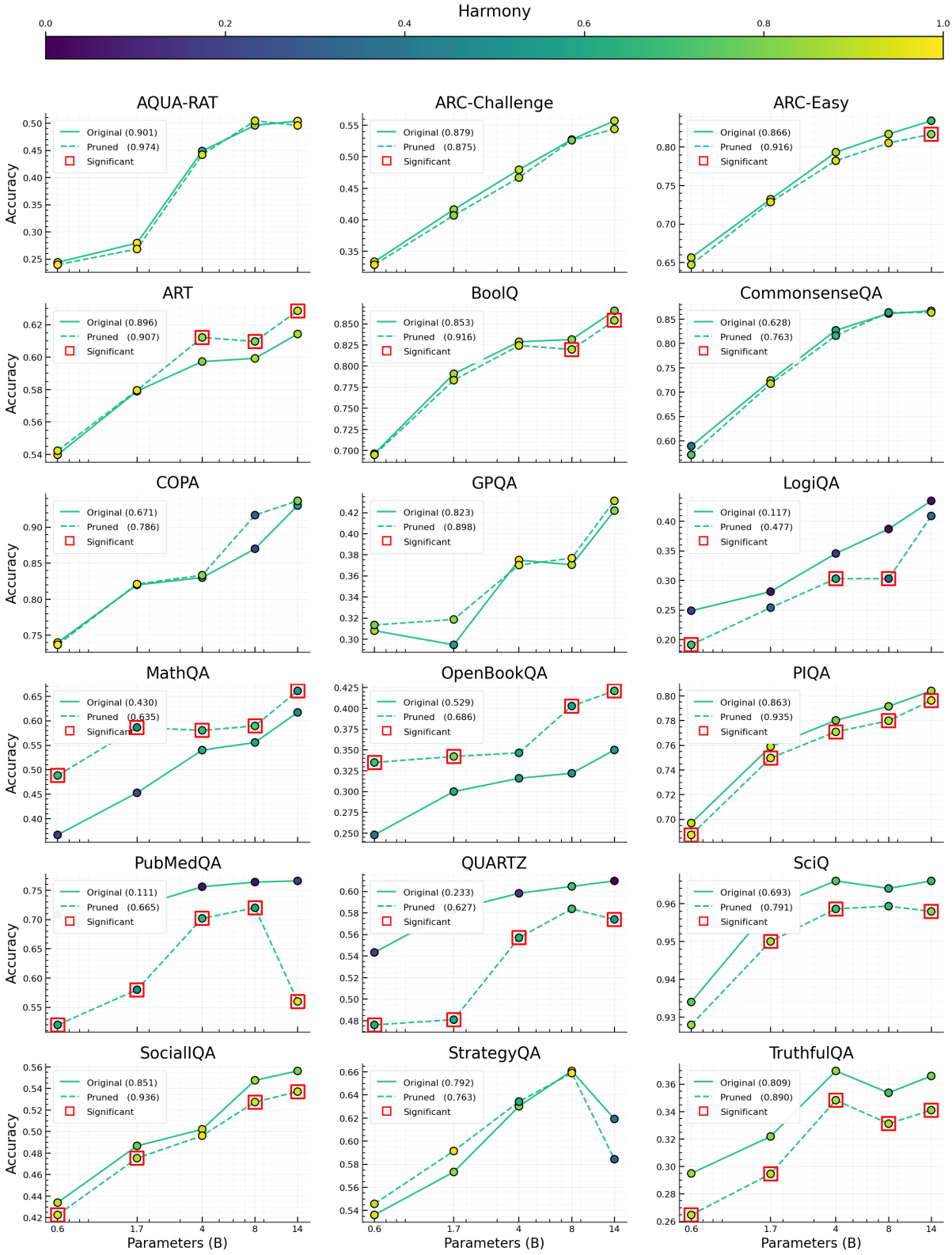


Figure 19: **Full results of balancing benchmarks via pruning in Qwen3 model family.** We remove overly similar items with a pruning rate inversely proportional to HARMONY, which consistently improves HARMONY. We find that aggregate scores often change statistically significantly on less harmonious benchmarks, whereas they remain more stable on more harmonious benchmarks.

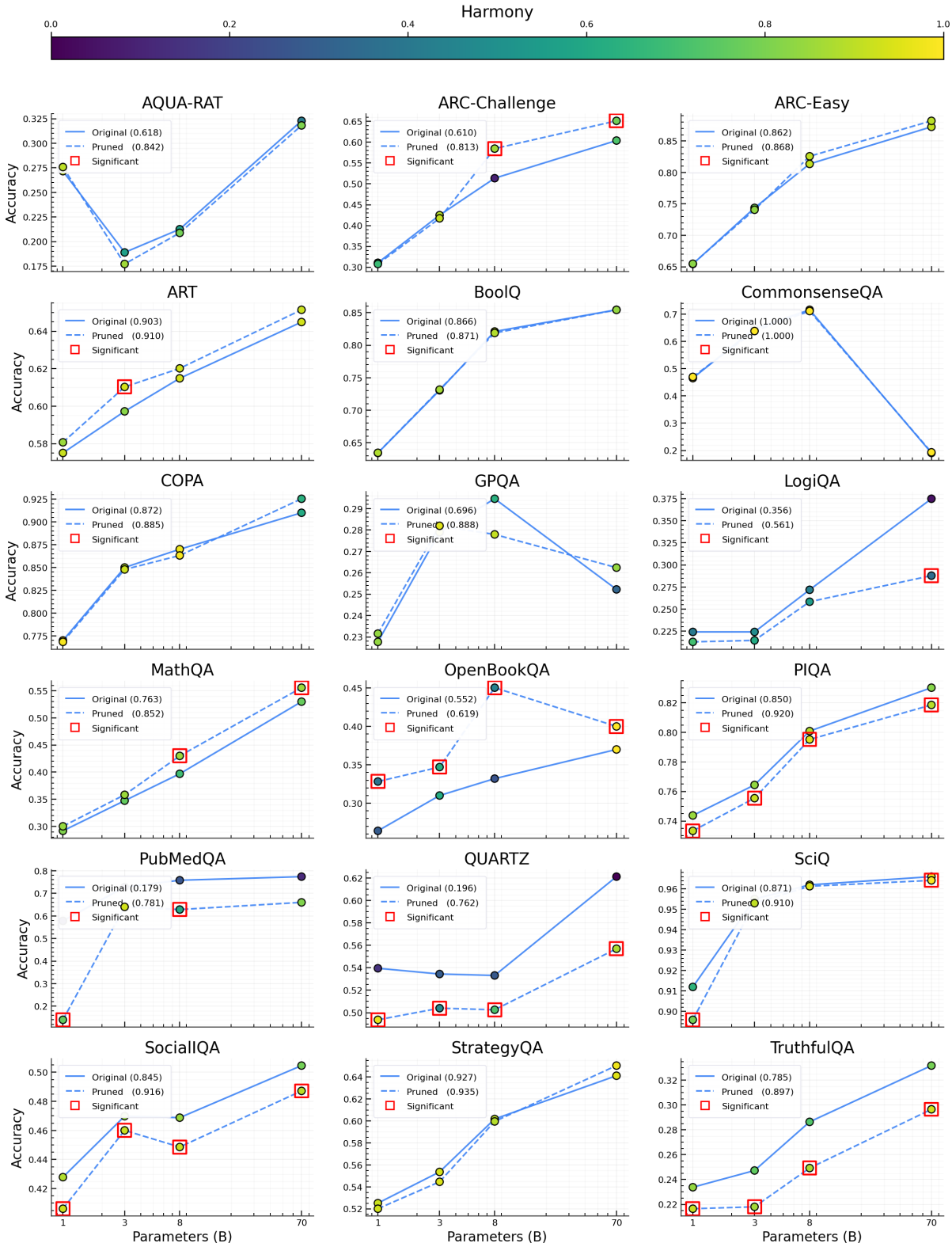


Figure 20: **Full results of balancing benchmarks via pruning in Llama 3 model family.** We remove overly similar items with a pruning rate inversely proportional to HARMONY, which consistently improves HARMONY. We find that aggregate scores often change statistically significantly on less harmonious benchmarks, whereas they remain more stable on more harmonious benchmarks.

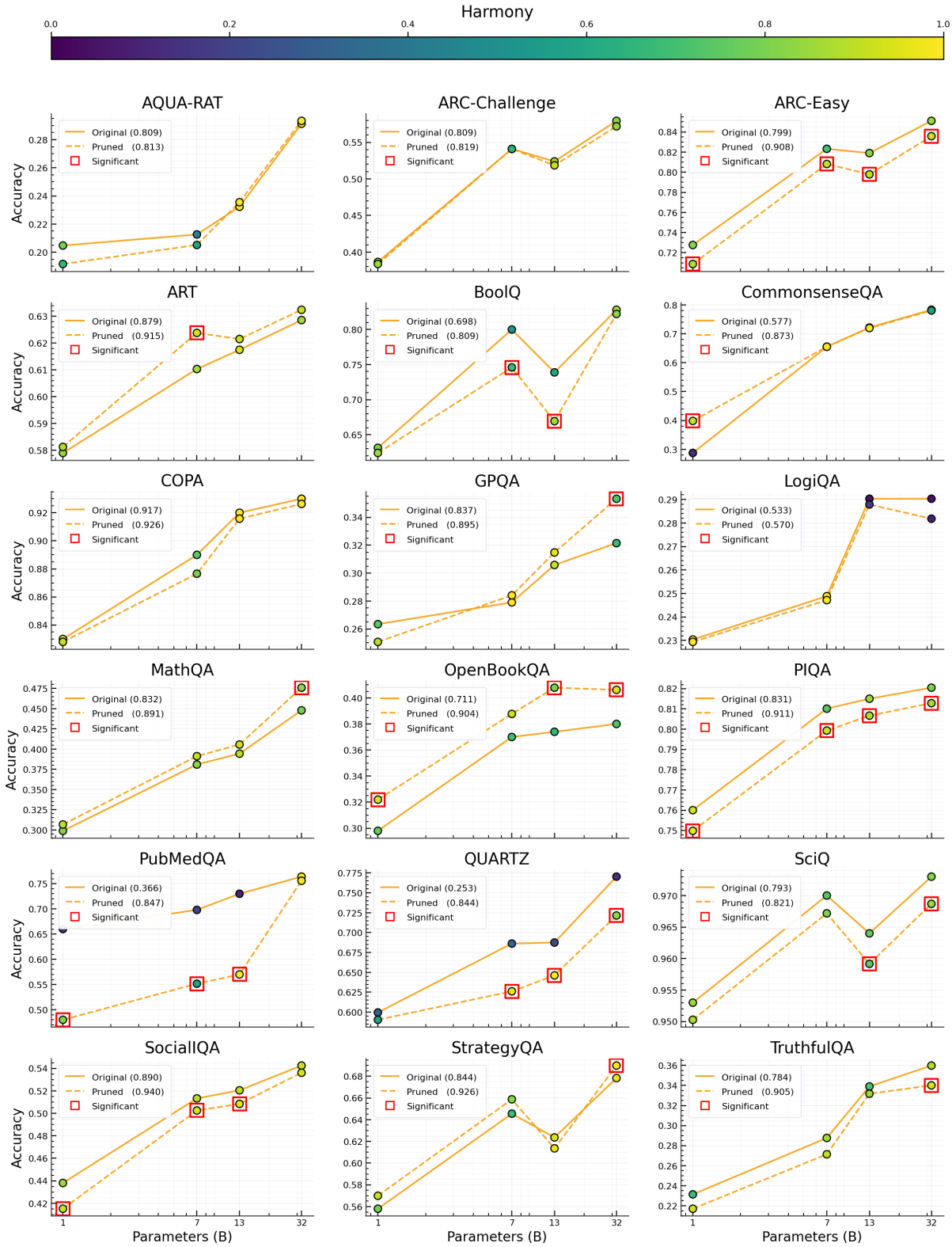


Figure 21: **Full results of balancing benchmarks via pruning in Olmo 2 model family.** We remove overly similar items with a pruning rate inversely proportional to HARMONY, which consistently improves HARMONY. We find that aggregate scores often change statistically significantly on less harmonious benchmarks, whereas they remain more stable on more harmonious benchmarks.

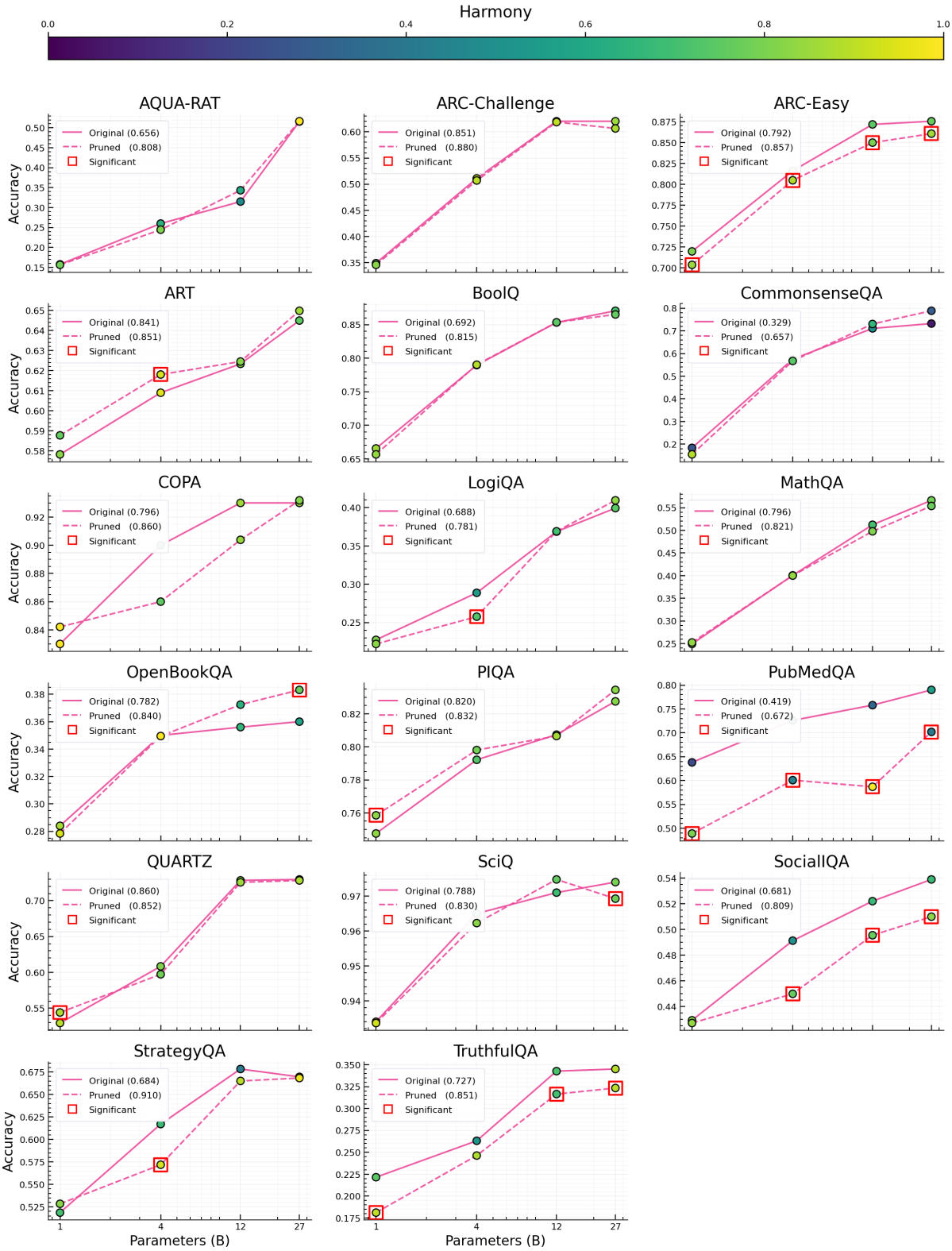


Figure 22: **Full results of balancing benchmarks via pruning in Gemma 3 model family.** We remove overly similar items with a pruning rate inversely proportional to HARMONY, which consistently improves HARMONY. We find that aggregate scores often change statistically significantly on less harmonious benchmarks, whereas they remain more stable on more harmonious benchmarks.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	47.2	68.1	82.6	86.8	**91.0**
Multicellular Biology (42.4%)	45.9	68.9	88.5	88.5	**93.4**
Evolutionary & Ecological Processes (25.7%)	51.4	59.5	86.5	**91.9**	**91.9**
Molecular & Cellular Biology (31.9%)	45.7	73.9	71.7	80.4	**87.0**
HARMONY	0.951	0.918	0.757	0.784	0.859

Table 10: Multi-dimensional evaluation results of the Qwen3 model family in MMLU College Biology. **Bold** implies the best performance.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	24.5	34.3	58.8	57.8	**69.6**
Quantum Mechanics Principles and Applications (22.5%)	56.5	34.8	65.2	69.6	**73.9**
Thermodynamics (8.8%)	0.0	55.6	66.7	55.6	**77.8**
Special Relativity Concepts (13.7%)	21.4	14.3	**71.4**	50.0	57.1
Classical Physics Principles and Relationships (20.6%)	9.5	38.1	47.6	52.4	**61.9**
Physics Phenomena and Application (16.7%)	0.0	11.8	11.8	17.6	**58.8**
Electromagnetism (5.9%)	16.7	16.7	**83.3**	**83.3**	66.7
Solid State Physics (11.8%)	50.0	75.0	**100.0**	**100.0**	**100.0**
HARMONY	0.686	0.712	0.765	0.815	0.789

Table 11: Multi-dimensional evaluation results of the Qwen3 model family in MMLU College Physics. **Bold** implies the best performance.

as reasoning (Bisk et al., 2019; Sap et al., 2019b; Zellers et al., 2019; Liu et al., 2020), world knowledge (Mihaylov et al., 2018b; Hendrycks et al., 2020; et al., 2023), and truthfulness (Lin et al., 2022; Khatun and Brown, 2024). Beyond individual benchmarks, holistic frameworks have emerged to offer a more comprehensive assessment of model performance (Liang et al., 2023; Chiang et al., 2024; Gao et al., 2024; Fourrier et al., 2023; et al., 2023). Reciprocally, understanding and improving current benchmarks have been equally important. MMLU Pro (Wang et al., 2024b) and Big-Bench-Hard (Suzgun et al., 2022) address benchmark saturation by constructing more challenging variants of MMLU (Hendrycks et al., 2020) and Big-Bench (bench authors, 2023) respectively. As top models approach ceiling effects on narrow probes, evaluation has shifted toward complex end-to-end tasks and composite suites. HLE and ARC-AGI assess multi-step reasoning, tool use, and robustness across domains (Phan et al., 2025; Chollet et al., 2025). Execution-grounded tasks such as SWE-bench measure real-world software problems and end-to-end correctness (Jimenez et al., 2024). Competitive exams like AIME and IMO, and professional exams such as the bar, push systems toward expert-level competence. Another recent practice is evaluation with online leaderboards, which use hidden test sets, fixed prompts, and compute disclosures in order to support fair comparison and consistent progress tracking (Chiang et al., 2024). Yet, these advances rest on a common premise that benchmarks reliably evaluate models on their stated domains. We audit this premise by testing whether benchmarks provide balanced coverage and promote comparable performance across subdomains.

P Hierarchical Labeling for Multi-Dimensional Evaluation

We build a tree benchmark over questions, then assign concise, human-readable labels to every node. Leaves summarize the shared evaluation focus of their questions, while internal nodes summarize their children.

To build the tree, we recursively induce partitions as discussed in §2.3, starting from the root (i.e., the entire benchmark) and ending at leaves (i.e., the clusters that do not admit a valid partition). For labeling leaves, we gather brief question annotations within a leaf and ask a model for one specific noun-phrase label. For labeling internal nodes of the tree, we pass the child labels to the model and ask for a slightly more abstract label that still captures the shared theme. Therefore, this procedure yields a bottom-up label propagation from leaves to internal nodes then to the root.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	60.7	72.2	80.5	83.5	84.2
Geology and Earth Sciences (15.0%)	64.4	73.4	83.8	83.8	83.8
Scientific Principles and Processes (32.6%)	56.3	70.1	79.9	83.5	84.3
Biological Processes and Concepts (24.1%)	65.4	74.0	78.8	82.9	83.6
Physics Principles in Engineering and Science (9.4%)	64.3	75.9	82.1	86.6	86.6
Environmental and Energy Assessment (6.5%)	60.4	63.6	78.6	81.2	79.9
Fundamental Concepts in Astronomy (7.6%)	56.1	76.7	80.6	83.3	87.2
Fundamentals of Chemical and Material Properties (4.8%)	56.1	71.9	81.6	84.2	84.2

Table 12: Multi-dimensional evaluation results for the Qwen3 model family on ARC-Easy.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	64.1	77.5	85.0	86.6	89.3
Product Composition, Properties, and Standards (8.8%)	66.8	80.6	83.7	86.9	88.9
Geographic, Operational, and Temporal Analysis (11.0%)	69.3	77.3	85.3	88.4	91.7
Media Standards and Analysis (24.8%)	61.3	77.7	86.8	88.7	91.0
Scientific and Analytical Principles (9.0%)	67.9	81.2	83.3	84.6	89.1
Sports History and Regulations (10.1%)	63.5	75.1	82.1	84.5	87.2
Governmental Laws and Regulations (10.6%)	64.1	72.8	80.3	84.6	86.1
Human Biology and Medical Science (6.5%)	60.6	81.2	87.8	85.4	87.8
Economic Systems (8.3%)	61.6	75.3	85.2	85.6	87.5
Sociocultural, Geopolitical, and Linguistic Analysis (7.1%)	63.9	76.8	88.0	86.7	89.7
Fictional Narrative Analysis and Elements (3.8%)	64.8	80.8	89.6	88.0	93.6

Table 13: Evaluation results for the Qwen3 model family on BoolQ.

We use `Gemini-2.0-flash` to annotate individual questions, assign each leaf a label from its question annotations, and propagate labels upward by aggregating child labels.

Q Example Clusters from Multi-Dimensional Evaluation

We present qualitative examples of the clusters produced by our hierarchical spectral clustering method, which groups questions using predictive similarity. Tables 28 and 29 illustrate representative clusters for ARC-Easy and BoolQ, respectively. Notably, successful hierarchical clustering requires coherence at each depth level, and these examples show that our method produces consistently coherent clusters even at lower depths, which is a particularly stringent and informative criterion.

R The Use of Large Language Models (LLMs)

We use large language models (LLMs) in two limited ways in this work. First, we use LLMs for light polishing of author-written text, including grammar, wording, and clarity edits. Second, we leverage LLMs to generate candidate synthetic data for RedundantQA. In particular, as described in detail in Appendix A, we begin with author-written seed examples and query `Gemini-2.0-flash` to expand them into candidate question sets, followed by automated consistency checks and manual review before inclusion in the final dataset.

All substantive content, methodological decisions, validation procedures, results, and conclusions are authored and verified by the authors, who take full responsibility for the paper's contents, including the synthetic data construction process and any text edited with LLM assistance.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	37.8	56.3	62.2	71.1	80.7
General Human Anatomy, Physiology, and Terminology (31.9%)	51.2	67.4	79.1	76.7	86.0
Head and Neck Anatomy (21.5%)	20.7	34.5	34.5	62.1	82.8
Skeletal Development and Anatomy (14.1%)	26.3	42.1	47.4	52.6	57.9
Neurological Disorders (10.4%)	21.4	50.0	50.0	71.4	78.6
Bone Anatomy and Terminology (3.0%)	0.0	50.0	50.0	50.0	50.0
Anatomy of Circulatory System (2.2%)	66.7	100.0	100.0	100.0	100.0
Developmental Structures (6.7%)	22.2	44.4	55.6	77.8	88.9
Nephrology (10.4%)	78.6	92.9	100.0	92.9	92.9

Table 14: Evaluation results for the Qwen3 model family on MMLU Anatomy.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	47.2	68.1	82.6	86.8	91.0
Multicellular Biology (42.4%)	45.9	68.9	88.5	88.5	93.4
Evolutionary and Ecological Processes (25.7%)	51.4	59.5	86.5	91.9	91.9
Molecular and Cellular Biology (31.9%)	45.7	73.9	71.7	80.4	87.0

Table 15: Evaluation results for the Qwen3 model family on MMLU College Biology.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	28.0	41.0	66.0	72.0	68.0
Theoretical Foundations of Computation (52.0%)	26.9	40.4	59.6	69.2	65.4
Computer Architecture and Optimization (7.0%)	28.6	42.9	71.4	57.1	57.1
Operating Systems (10.0%)	50.0	80.0	90.0	90.0	80.0
Network Layer Protocols and Technologies (5.0%)	60.0	60.0	80.0	100.0	100.0
Data Processing (12.0%)	8.3	8.3	66.7	50.0	66.7
Sorting Algorithms (4.0%)	25.0	25.0	75.0	100.0	100.0
Graph Algorithms and Data Structures (10.0%)	20.0	40.0	60.0	80.0	50.0

Table 16: Evaluation results for the Qwen3 model family on MMLU College Computer Science.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	31.0	39.0	55.0	59.0	68.0
Advanced Real Analysis (19.0%)	26.3	26.3	42.1	52.6	52.6
Abstract Algebra (11.0%)	27.3	45.5	90.9	63.6	81.8
Probability (7.0%)	28.6	71.4	28.6	57.1	57.1
Properties of Mathematical Operations and Functions (5.0%)	20.0	20.0	40.0	80.0	80.0
Advanced Mathematical Concepts and Applications (16.0%)	31.3	37.5	75.0	56.3	75.0
Mathematical Modeling and Algorithms (10.0%)	30.0	50.0	60.0	60.0	70.0
Multivariable Calculus (27.0%)	25.9	37.0	40.7	59.3	63.0
Mathematical Optimization Methods (5.0%)	100.0	40.0	80.0	60.0	100.0

Table 17: Evaluation results for the Qwen3 model family on MMLU College Mathematics.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	24.5	34.3	58.8	57.8	69.6
Quantum Mechanics Principles and Applications (22.5%)	56.5	34.8	65.2	69.6	73.9
Thermodynamics (8.8%)	0.0	55.6	66.7	55.6	77.8
Special Relativity Concepts (13.7%)	21.4	14.3	71.4	50.0	57.1
Classical Physics Principles and Relationships (20.6%)	9.5	38.1	47.6	52.4	61.9
Physics Phenomena and Applications (16.7%)	0.0	11.8	11.8	17.6	58.8
Electromagnetism (5.9%)	16.7	16.7	83.3	83.3	66.7
Solid State Physics Concepts (11.8%)	50.0	75.0	100.0	100.0	100.0

Table 18: Evaluation results for the Qwen3 model family on MMLU College Physics.

	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
Overall	52.5	66.2	83.3	88.7	91.7
US Sociopolitical Ideologies, Movements, and Issues (46.1%)	52.1	64.9	83.0	90.4	95.7
Progressive Era Economic and Social Initiatives (3.9%)	75.0	87.5	100.0	100.0	100.0
United States Governance and Politics (37.3%)	46.1	61.8	81.6	85.5	88.2
Ideological and Territorial Expansion in the Americas (9.8%)	70.0	80.0	90.0	95.0	85.0
American History Eras (2.9%)	50.0	66.7	66.7	66.7	83.3

Table 19: Evaluation results for the Qwen3 model family on MMLU High School U.S. History.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	72.0	81.6	87.2	87.5
Geology and Earth Sciences (15.0%)	74.5	84.6	87.7	89.4
Scientific Principles and Processes (32.6%)	71.2	79.7	86.5	87.0
Biological Processes and Concepts (24.1%)	74.3	83.0	88.5	87.8
Physics Principles in Engineering and Science (9.4%)	70.5	83.0	88.8	88.8
Environmental and Energy Assessment (6.5%)	68.2	80.5	83.1	84.4
Fundamental Concepts in Astronomy (7.6%)	72.8	79.4	86.7	87.2
Fundamentals of Chemical and Material Properties (4.8%)	64.0	79.8	86.8	86.8

Table 20: Multi-dimensional evaluation results for the Gemma 3 model family on ARC-Easy.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	66.5	79.0	85.3	87.1
Product Composition, Properties, and Standards (8.8%)	67.5	76.1	82.4	88.6
Geographic, Operational, and Temporal Analysis (11.0%)	67.6	83.7	84.8	87.3
Media Standards and Analysis (24.8%)	66.7	80.6	88.5	89.6
Scientific and Analytical Principles (9.0%)	65.2	75.1	82.9	86.0
Sports History and Regulations (10.1%)	63.2	78.7	83.0	86.0
Governmental Laws and Regulations (10.6%)	64.1	75.7	80.3	82.0
Human Biology and Medical Science (6.5%)	77.0	82.6	86.9	89.2
Economic Systems (8.3%)	66.1	77.1	85.6	85.6
Sociocultural, Geopolitical, and Linguistic Analysis (7.1%)	61.8	78.1	87.6	84.1
Fictional Narrative Analysis and Elements (3.8%)	71.2	79.2	91.2	90.4

Table 21: Multi-dimensional evaluation results for the Gemma 3 model family on BoolQ.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	25.9	61.5	70.4	70.4
General Human Anatomy, Physiology, and Terminology (31.9%)	27.9	69.8	83.7	79.1
Head and Neck Anatomy (21.5%)	27.6	41.4	55.2	48.3
Skeletal Development and Anatomy (14.1%)	21.1	52.6	52.6	57.9
Neurological Disorders (10.4%)	42.9	64.3	57.1	71.4
Bone Anatomy and Terminology (3.0%)	50.0	50.0	50.0	50.0
Anatomy of Circulatory System (2.2%)	0.0	100.0	100.0	100.0
Developmental Structures (6.7%)	11.1	44.4	77.8	88.9
Nephrology (10.4%)	14.3	92.9	92.9	92.9

Table 22: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU Anatomy.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Multicellular Biology (42.4%)	27.9	68.9	98.4	93.4
Evolutionary and Ecological Processes (25.7%)	21.6	67.6	86.5	86.5
Molecular and Cellular Biology (31.9%)	21.7	65.2	87.0	91.3

Table 23: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU College Biology.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	30.0	48.0	57.0	63.0
Theoretical Foundations of Computation (52.0%)	36.5	38.5	51.9	61.5
Computer Architecture and Optimization (7.0%)	28.6	57.1	71.4	57.1
Operating Systems (10.0%)	20.0	60.0	80.0	80.0
Network Layer Protocols and Technologies (5.0%)	20.0	60.0	100.0	100.0
Data Processing (12.0%)	25.0	33.3	41.7	41.7
Sorting Algorithms (4.0%)	0.0	100.0	100.0	75.0
Graph Algorithms and Data Structures (10.0%)	30.0	70.0	30.0	60.0

Table 24: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU College Computer Science.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	33.0	41.0	50.0	58.0
Advanced Real Analysis (19.0%)	57.9	52.6	47.4	52.6
Abstract Algebra (11.0%)	27.3	72.7	63.6	54.5
Probability (7.0%)	14.3	57.1	42.9	57.1
Properties of Mathematical Operations and Functions (5.0%)	60.0	40.0	40.0	80.0
Advanced Mathematical Concepts and Applications (16.0%)	25.0	31.3	56.3	68.8
Mathematical Modeling and Algorithms (10.0%)	20.0	30.0	50.0	60.0
Multivariable Calculus (27.0%)	29.6	25.9	40.7	55.6
Mathematical Optimization Methods (5.0%)	20.0	40.0	80.0	40.0

Table 25: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU College Mathematics.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	20.6	41.2	52.9	63.7
Quantum Mechanics Principles and Applications (22.5%)	8.7	39.1	43.5	73.9
Thermodynamics (8.8%)	11.1	22.2	55.6	55.6
Special Relativity Concepts (13.7%)	28.6	35.7	42.9	50.0
Classical Physics Principles and Relationships (20.6%)	42.9	28.6	57.1	66.7
Physics Phenomena and Applications (16.7%)	5.9	41.2	29.4	41.2
Electromagnetism (5.9%)	0.0	66.7	83.3	50.0
Solid State Physics Concepts (11.8%)	33.3	75.0	91.7	100.0

Table 26: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU College Physics.

	Gemma 3 1B	Gemma 3 4B	Gemma 3 12B	Gemma 3 27B
Overall	26.0	75.5	88.2	91.2
US Sociopolitical Ideologies, Movements, and Issues (46.1%)	26.6	80.9	87.2	92.6
Progressive Era Economic and Social Initiatives (3.9%)	25.0	100.0	100.0	87.5
United States Governance and Politics (37.3%)	26.3	67.1	89.5	90.8
Ideological and Territorial Expansion in the Americas (9.8%)	20.0	75.0	85.0	90.0
American History Eras (2.9%)	33.3	66.7	83.3	83.3

Table 27: Multi-dimensional evaluation results for the Gemma 3 model family on MMLU High School U.S. History.

ID	Depth	Topic	Questions
2139	5	Reproductive Systems	1) Which is the greatest benefit of sexual reproduction? 2) Which of these is the best example of sexual reproduction? 3) Which is an example of asexual reproduction? 4) Which statement describes a characteristic of both sexual and asexual reproduction?
1244	5	Photosynthesis	1) In what form do plants store the energy produced from sunlight? 2) Which of the following best describes how plants use the energy they receive from sunlight? 3) Which of the following processes makes it possible for plants to use energy from sunlight to produce their own food? 4) Which form of energy do plants need to capture in order to perform photosynthesis?
163	3	Waves	1) What causes sound? 2) A substance that carries sound waves is called 3) An echo is a sound wave that has been 4) Sound travels as a
131	2	Landform Processes	1) A beach is formed when sediment is deposited along a shoreline. What would most likely happen if rivers that empty into the ocean were dammed? 2) As a river enters a larger body of water, sediments are deposited over a wide area. Which of these landforms is likely to be formed at the site of deposition? 3) Deposition of sediment will most likely form a 4) Which statement describes the formation of a delta?
2206	3	Laboratory Equipment	1) Which tool is best used to observe a soil sample? 2) Which tool would be most useful for observing the details of an insect's wings? 3) Which tools are used to determine the boiling point of water? 4) Which device is used to determine the volume of a liquid?

Table 28: Example ARC-Easy clusters with their associated ID, depth, and topic.

ID	Depth	Topic	Questions
2818	2	Locations	1) is st augustine the oldest city in florida 2) is san juan puerto rico in the caribbean 3) is st augustine florida the oldest city in america 4) is there an amtrak station in pensacola florida 5) is key west part of the united states 6) is riviera maya on the gulf of mexico 7) is panama city beach on the gulf of mexico
3026	2	Species Hybridization	1) is artificial selection and selective breeding the same thing 2) can a grizzly bear mate with a polar bear 3) can a horse and a donkey have a baby 4) can a polar bear and a grizzly mate 5) is it possible for a zebra and a horse to interbreed
1751	3	Sports	1) did kaka play in the 2002 world cup 2) is the uefa champions league final one game 3) did alphonse areola play in world cup 2018 4) has a keeper ever won the ballon d'or 5) has christiano ronaldo ever won the world cup
1149	3	Breaking Bad	1) will there be a season six of breaking bad 2) is better call saul set after breaking bad 3) was better call saul filmed before breaking bad 4) is there a spin off from breaking bad
497	2	Constitutional Rights	1) is the first amendment in the bill of rights 2) was the right to bear arms in the original constitution 3) does the us constitution protect the right to privacy 4) do corporations have the same free speech rights as persons 5) are fighting words protected under freedom of speech 6) does the first amendment separate church and state 7) does the right to privacy exist in the constitution 8) is there a limit on freedom of speech 9) is the second amendment part of the constitution

Table 29: Example BoolQ clusters with their associated ID, depth, and topic.