# BenchMetrics: A Framework for Evaluating LLM Benchmarks

**Anonymous ACL submission**

## Abstract

The rapid proliferation of benchmarks for evaluating large language models (LLMs) has created an urgent need for systematic methods to assess benchmark quality itself. We propose BENCHMETRICS, a framework that quantifies benchmark quality along three complementary axes: (1) **Cross-Benchmark Ranking Consistency**, measuring whether a benchmark produces model rankings aligned with peer benchmarks; (2) **Discriminability Score**, quantifying a benchmark's ability to differentiate between models; and (3) **Capability Alignment Deviation**, identifying problematic instances where stronger models fail but weaker models succeed *within the same model family*. Across 15 benchmarks spanning mathematics, reasoning, and knowledge domains, evaluated on 11 instruction-tuned LLMs from four model families (plus 3 base LLMs reserved for held-out validation), BENCHMETRICS reveals significant quality variations among existing benchmarks and shows that selective benchmark construction guided by these metrics achieves comparable evaluation fidelity with substantially reduced test sets.

## 1 Introduction

The evaluation of large language models (LLMs) has become increasingly important as these models are deployed across diverse real-world applications. Benchmarks serve as the primary instruments for measuring model capabilities, guiding both research directions and practical deployment decisions. However, the explosive growth in the number of benchmarks, with hundreds now available across domains such as mathematics, reasoning, instruction following, and knowledge understanding, raises a fundamental question: *How do we know if a benchmark itself is good?*

Although benchmarks serve an important role in the field, surprisingly little attention has been paid to the benchmark quality. Current practice often treats benchmarks as ground truth without questioning their reliability or validity. This oversight can lead to several problems, as illustrated in Figure 1: (1) **Ranking Inconsistency**: different benchmarks may produce conflicting model rankings, making it unclear which benchmark to trust; (2) **Low Discriminative Power**: some benchmarks fail to differentiate between models of varying capabilities, clustering all models within a narrow performance range; and (3) **Rank-Inconsistent Items**: individual test instances may exhibit counter-intuitive behavior where stronger models fail but weaker models succeed.

For example, if Benchmark A ranks Model X above Model Y while several peer benchmarks consistently disagree, should we trust A; and if a benchmark assigns near-identical scores to a frontier model and a much smaller one, does that reflect similar capabilities or simply a lack of discriminative power?

To address these challenges, we propose BENCHMETRICS, a framework that scores benchmark quality along three complementary metrics (Figure 1, bottom row): **Cross-Benchmark Ranking Consistency (CBRC)** measures the ranking correlation of a benchmark with other benchmarks in the same domain; **Discriminability Score (DS)** quantifies how well a benchmark separates models of varying capabilities; and **Capability Alignment Deviation (CAD)** penalizes counter-intuitive instances where weaker models outperform stronger ones within the same family, ensuring hierarchical consistency.

We conduct comprehensive experiments across 15 widely-used benchmarks spanning three major domains (mathematics, general reasoning, and knowledge & understanding), evaluating 11 instruction-tuned LLMs across four model families with clear capability hierarchies and an additional 3 base LLMs reserved for held-out validation (14 models total). Our analysis reveals substantial

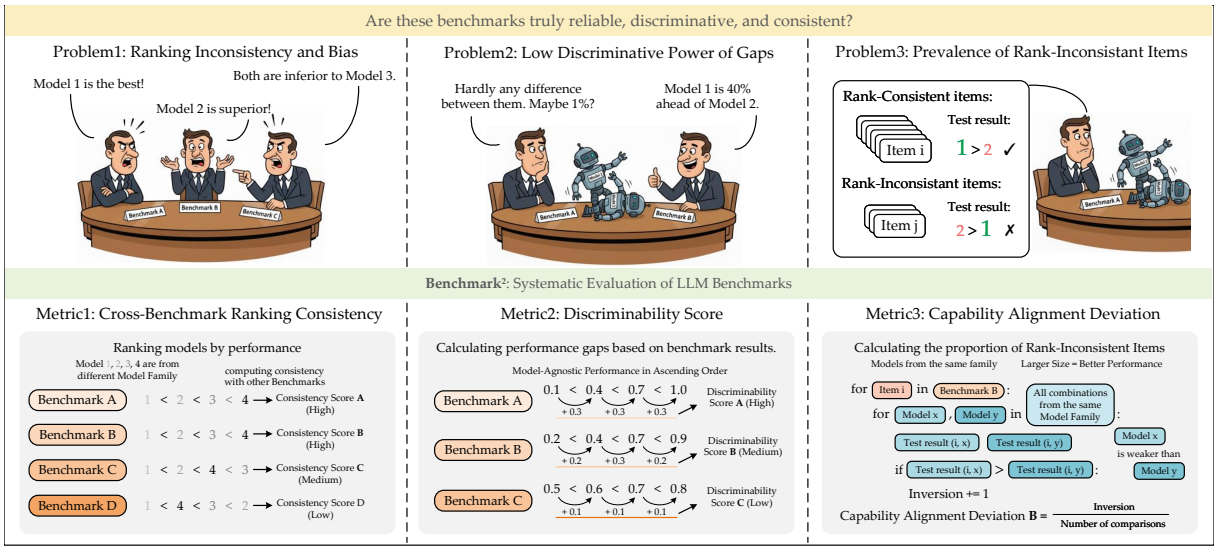


Figure 1: Overview of BENCHMETRICS framework. **Top row**: Three key problems with existing LLM benchmarks: ranking inconsistency across benchmarks, low discriminative power of performance gaps, and prevalence of rank-inconsistent test items. **Bottom row**: Our three complementary metrics addressing each problem: Cross-Benchmark Ranking Consistency (CBRC) measures alignment with peer benchmarks, Discriminability Score (DS) quantifies performance gap magnitudes, and Capability Alignment Deviation (CAD) identifies items violating expected capability hierarchies within model families.

quality variations among existing benchmarks and identifies specific characteristics that distinguish high-quality benchmarks from problematic ones.

Beyond benchmark quality assessment, we demonstrate a practical application of our framework: selective benchmark construction. By identifying high-quality test instances based on our metrics, we construct reduced benchmark versions that achieve comparable evaluation performance to full benchmarks while providing greater efficiency.

Our contributions are summarized as follows:

- We formalize the problem of benchmark quality assessment and propose three complementary metrics that capture different aspects of benchmark reliability.
- We conduct the first large-scale systematic evaluation of benchmark quality across 15 benchmarks and 11 models spanning four families, providing empirical insights into the state of LLM evaluation.
- We show that filtering instances via quality metrics achieves comparable evaluation performance using only 35% of the original data.

## 2 Related Work

### 2.1 LLM Benchmarks

The landscape of LLM evaluation has expanded dramatically. General-purpose benchmarks such as MMLU (Hendrycks et al., 2021a), BBH (Suzgun et al., 2023), and ARC (Clark et al., 2018) measure broad capabilities across multiple domains. Domain-specific benchmarks have emerged for mathematics (MATH-500, AIME, OlympiadBench), reasoning and comprehension (DROP (Dua et al., 2019), CommonsenseQA (Talmor et al., 2019)), and knowledge understanding (IFEval(Zhou et al., 2023), SuperGPQA (M-A-P Team et al., 2025)). More challenging benchmarks have been introduced to address ceiling effects, including OmniMath (Gao et al., 2024) for advanced mathematics. However, this proliferation has occurred largely without systematic quality assessment.

### 2.2 Benchmark Quality Analysis

Several studies have examined issues with existing benchmarks. Bowman and Dahl (2021) discussed dangers of benchmark-driven research and the need for more robust evaluation practices. Data contamination, where LLMs inadvertently encounter test data during training, has been identified as a significant concern (Xu et al., 2024; Sainz et al., 2023). Benchmark saturation has motivated dynamic benchmarks (Kiela et al., 2021). Research on evaluation methodology has addressed statistical significance in model comparisons (Dror et al., 2018) and limitations of single-number metrics

(Ethayarajh and Jurafsky, 2020). Liang et al. (2023) introduced HELM for holistic evaluation across multiple dimensions. Psychometric tools such as Item Response Theory have also been adapted for analyzing test items in classical NLP settings, though they typically assume a single latent skill and were not designed for the heterogeneous training recipes seen across modern LLM families.

Our work complements these efforts by targeting a different question: rather than running an evaluation (HELM), refreshing benchmarks (Dynabench), or estimating item difficulty under a unidimensional latent trait (IRT), we provide *quantitative metrics for assessing benchmark quality itself*. Crucially, our metrics exploit two structural properties of the LLM ecosystem that prior tools do not leverage: (i) the availability of multiple peer benchmarks per domain (used by CBRC), and (ii) within-family parameter-count hierarchies (used by CAD), which give a reliable ordering even when cross-family comparisons are not. These design choices make BENCHMETRICS distinct from existing meta-evaluation approaches.

## 3 Methodology

In this section, we formally introduce the three quality metrics that constitute BENCHMETRICS. Consider a set of benchmarks $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ and a set of candidate models $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$. Let $s_{ij}$ denote the performance score of model $M_j$ evaluated on benchmark $B_i$. Based on this formulation, we propose three complementary approaches to assess benchmark quality:

### 3.1 Cross-Benchmark Ranking Consistency

Cross-Benchmark Ranking Consistency (CBRC) evaluates whether a benchmark's ranking corroborates with others in the same domain. The underlying rationale is that effective benchmarks measuring similar capabilities should produce highly correlated model rankings.

**Definition.** For a benchmark $B_i$, we compute its ranking consistency as the average Kendall's $\tau$ correlation with other benchmarks in the same domain:

$$\text{CBRC}(B_i) = \frac{1}{n-1} \sum_{j \neq i} \tau(r_i, r_j) \qquad (1)$$

where $r_i$ denotes the ranking of models induced by benchmark $B_i$, and $\tau(\cdot, \cdot)$ is Kendall's tau correlation coefficient.

**Interpretation.** CBRC values range from -1 to 1, where 1 indicates perfect agreement with other benchmarks, 0 indicates no correlation, and negative values indicate inverse rankings. We consider CBRC $> 0.7$ as indicating high consistency, $0.4$–$0.7$ as moderate, and $< 0.4$ as low.

**What CBRC does and does not measure.** CBRC captures *external agreement* with peer benchmarks; it does not directly measure intrinsic difficulty, novelty, or coverage. A benchmark that legitimately stresses a capability under-represented elsewhere may correctly receive a lower CBRC, and a high CBRC could in principle reflect shared training-data contamination or overlapping formats rather than genuine quality. For this reason, CBRC is intended to be used alongside the reference-free metrics DS and CAD: the latter two do not rely on any external benchmarks, and together the three metrics provide complementary evidence about a benchmark's quality.

### 3.2 Discriminability Score

A high-quality benchmark should effectively differentiate between models of varying capabilities. If all models achieve similar scores regardless of their actual ability differences, the benchmark provides limited useful information.

**Definition.** We define the Discriminability Score (DS) based on the normalized score spread and the statistical significance of pairwise differences:

$$\text{DS}(B_i) = \frac{\sigma_i}{\bar{s}_i} \cdot \sqrt{\frac{\sum_{j<k} \mathbf{1}[|s_{ij} - s_{ik}| > \epsilon]}{m(m-1)/2}} \qquad (2)$$

where $\sigma_i$ is the standard deviation of scores on benchmark $B_i$, $\bar{s}_i$ is the mean score, and the second term represents the proportion of model pairs with practically significant differences (we set $\epsilon = 0.02$ as the minimum meaningful difference).

**Interpretation.** Higher DS values indicate better discriminability. We empirically find that benchmarks with DS $> 0.4$ provide good differentiation, while those with DS $< 0.2$ offer minimal discrimination between models.

### 3.3 Capability Alignment Deviation

This metric operates at the instance level, identifying individual test questions that may be problematic. The key insight is that if a question is well-designed, stronger models should generally outperform weaker models on it, maintaining alignment with expected capability hierarchies.

**Model Family Hierarchy.** Rather than establishing a global ordering across all models, which can be unreliable due to different training data and optimization objectives across model families, we leverage the natural capability hierarchy *within* model families. For a model family $\mathcal{F}$ (e.g., Qwen2.5-Instruct), we define an ordering based on parameter count:

$$\mathcal{F} = \{M_1 \succ M_2 \succ \cdots \succ M_k\} \tag{3}$$

where $M_1$ has the most parameters and $M_k$ has the fewest. This within-family ordering is more reliable than cross-family comparisons.

**Definition.** For a benchmark $B_i$, we first compute the raw inversion rate by aggregating inversions across all model families $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_F\}$:

$$\text{inv_rate}(B_i) = \frac{\sum_{f=1}^{F} \text{inv}_{\mathcal{F}_f}(B_i)}{\sum_{f=1}^{F} \text{comp}_{\mathcal{F}_f}(B_i)} \tag{4}$$

For each family $\mathcal{F} = \{M_1 \succ M_2 \succ \cdots \succ M_k\}$, an inversion on question $q$ occurs when a stronger model fails but a weaker model succeeds:

$$\text{inv}_{\mathcal{F}}(B_i) = \sum_{q \in Q_i} \sum_{j<l} \mathbf{1}[\neg c_{jq} \wedge c_{lq}] \tag{5}$$

where $Q_i$ is the set of questions in benchmark $B_i$, $c_{jq}$ indicates whether model $M_j$ correctly answers question $q$.

We then apply an exponential transformation to convert the inversion rate to a score where higher values indicate better alignment:

$$\text{CAD}(B_i) = e^{-\lambda \cdot \text{inv_rate}(B_i)} \tag{6}$$

where $\lambda > 0$ is a scaling parameter that controls the sensitivity of the transformation. In our experiments, we set $\lambda = 12$ based on empirical analysis to ensure meaningful differentiation across the observed range of inversion rates.

**Interpretation.** CAD ranges from 0 to 1, where 1 indicates perfect alignment (no inversions) and values approaching 0 indicate severe capability hierarchy violations. We consider CAD $> 0.6$ as indicating good quality, 0.4–0.6 as acceptable, and $< 0.4$ as indicating significant quality issues.

**Robustness to isolated and family-specific inversions.** Importantly, the inversion rate in Eq. (4) aggregates over *all* model families. A given item only contributes substantially to a benchmark's CAD penalty when multiple families exhibit inversions on it, i.e., when the inversion is *systematic* rather than family-specific. Phenomena such as "U-shaped" scaling, RLHF-induced quirks, or domain-specific peculiarities of a single family (e.g., a chat-tuned variant outperforming a larger pre-trained sibling on social tasks) do not produce inversions on the same item across all families simultaneously, and therefore have only a small effect on CAD. The transform mapping in App. B (Table 8) makes this concrete: a small fraction of isolated inversions (raw rate $\leq 0.02$) maps to CAD $\geq 0.79$, well within the "Excellent" range.

### 3.4 Stability Score

To assess the reliability of selective benchmark evaluation, we introduce the Stability Score, which measures the consistency of model rankings across multiple sampling iterations.

**Definition.** For a selective benchmark $B_s$ with selection ratio $r$, we perform $K$ bootstrap sampling iterations (we use $K = 100$). In each iteration $k$, we sample $r \cdot |B|$ instances and compute the resulting model ranking $r_k$. The Stability Score is defined as the average pairwise ranking correlation:

$$\text{Stability}(B_s) = \frac{2}{K(K-1)} \sum_{i<j} \tau(r_i, r_j) \tag{7}$$

where $\tau(\cdot, \cdot)$ is Kendall's tau correlation coefficient between rankings from different bootstrap samples.

**Interpretation.** Stability Score ranges from -1 to 1, where 1 indicates that the selective benchmark produces identical rankings regardless of which specific instances are sampled, and lower values indicate higher variance in rankings. We consider Stability $> 0.7$ as high, 0.5–0.7 as moderate, and $< 0.5$ as low.

### 3.5 Benchmark Quality Score

We also provide a combined score for overall assessment:

$$\begin{aligned}\text{BQS}(B_i) = \alpha \cdot \widetilde{\text{CBRC}}(B_i) + \beta \cdot \text{DS}(B_i) \\ + \gamma \cdot \text{CAD}(B_i)\end{aligned} \tag{8}$$

where $\widetilde{\text{CBRC}}$ denotes the normalized CBRC score, and $\alpha$, $\beta$, $\gamma$ are weighting parameters. Details on normalization and weight selection are provided in Appendix C.

## 4 Experimental Setup

### 4.1 Benchmarks

To demonstrate the broad applicability of our framework, we select a diverse set of 15 benchmarks across three major domains:

**Mathematics (5 benchmarks).** AIME 2024 (Mathematical Association of America, 2024a), OmniMath (Gao et al., 2024) for advanced mathematical problem solving, OlympiadBench (He et al., 2024), AMC (Mathematical Association of America, 2024b) and MATH-500 (Hendrycks et al., 2021b).

**General Reasoning (5 benchmarks).** BigBench Hard (BBH) (Suzgun et al., 2023) for challenging tasks, DROP (Dua et al., 2019) for reading comprehension, ARC (Clark et al., 2018) for scientific reasoning, CommonsenseQA (Talmor et al., 2019) for commonsense reasoning and SIQA (Sap et al., 2019) for social intelligence.

**Knowledge & Understanding (5 benchmarks).** SuperGPQA (M-A-P Team et al., 2025) for graduate-level reasoning, MMLU-Pro (Wang et al., 2024) , IFBench (Pyatkin et al., 2025) and IFEval (Zhou et al., 2023) for instruction following capabilities and EQ-Bench (Paech, 2023) for emotional intelligence.

### 4.2 Models

We evaluate 11 models across four families with clear capability hierarchies based on model size:

- **DeepSeek-R1-Distill-Qwen (DeepSeek-AI, 2025)**: 1.5B, 7B, 32B
- **Llama-3.1-Instruct (Grattafiori et al., 2024)**: 8B, 70B
- **Qwen2.5-Instruct (Yang et al., 2024)**: 1.5B, 7B, 72B
- **Qwen3 (Yang et al., 2025)**: 1.7B, 8B, 32B

**Selection Rationale.** To ensure architectural diversity and enable reliable CAD computation across multiple scales, we selected models from four distinct development lineages. This strategy mitigates the risk of family-specific bias while providing 1–3 comparison pairs within each family, yielding a total of 10 pairs per benchmark instance.

**Held-Out Validation.** To verify that our metrics generalize beyond the models used in their computation, we additionally evaluate on **Qwen2.5-Base** (1.5B, 7B, 32B), which share the same architecture as Qwen2.5-Instruct but were **not used** in metric computation and have fundamentally different training (no instruction tuning).

### 4.3 Evaluation Protocol

For each model-benchmark pair, we use standardized prompting templates following the original benchmark specifications where available. We use the default vLLM sampling configuration (temperature $= 0.7$, top-$p = 0.8$; see App. A, Table 6) and evaluate using exact match or execution-based metrics as appropriate for each benchmark. To control variance from sampling, all of our metrics are aggregated over multiple instance- and model-level comparisons; the bootstrap confidence intervals in App. E confirm that the resulting estimates are stable.

## 5 Results

### 5.1 Comprehensive Quality Analysis

Table 1 reveals distinct quality profiles across domains.

**Mathematics** exhibits the widest quality variation (BQS: 0.55–0.79). AIME 2024 achieves exceptional discriminability (DS = 0.74) and capability alignment (CAD = 0.85), while MATH-500 shows potential ceiling effects with low discriminability (DS = 0.16).

**General Reasoning** presents a quality-discriminability trade-off. ARC achieves the highest capability alignment (CAD = 0.87) but limited discriminability (DS = 0.11), whereas BBH maximizes discriminability (DS = 0.25) at the cost of alignment (CAD = 0.66). SIQA exhibits problematic quality across all model families (CAD = 0.23).

**Knowledge & Understanding** shows the most consistent quality profile (BQS: 0.51–0.58), with IFEval and SuperGPQA achieving strong cross-benchmark consistency (CBRC $\geq 0.75$).

Two patterns emerge: (1) high discriminability and high capability alignment rarely co-occur, since the most discriminative benchmarks tend to

| Benchmark | CBRC | | DS | | | CAD | | BQS |
|---|---|---|---|---|---|---|---|---|
| | Score | $\sigma$ | Score | Range | $\sigma$ | Score | $\sigma$ | |
| *Mathematics* | | | | | | | | |
| AIME 2024 | 0.52 | 0.10 | 0.74 | 0–53 | 0.22 | 0.85 | 0.13 | 0.79 |
| OmniMath | 0.76 | 0.13 | 0.79 | 0–62 | 0.27 | 0.61 | 0.22 | 0.75 |
| OlympiadBench | 0.75 | 0.12 | 0.76 | 0–64 | 0.32 | 0.61 | 0.23 | 0.73 |
| AMC 22-24 | 0.70 | 0.18 | 0.36 | 16–67 | 0.11 | 0.46 | 0.09 | 0.55 |
| MATH-500 | 0.70 | 0.18 | 0.16 | 49–87 | 0.07 | 0.62 | 0.10 | 0.55 |
| *General Reasoning* | | | | | | | | |
| ARC | 0.79 | 0.03 | 0.11 | 56–95 | 0.07 | 0.87 | 0.05 | 0.65 |
| BBH | 0.75 | 0.02 | 0.25 | 30–89 | 0.12 | 0.66 | 0.05 | 0.60 |
| DROP | 0.71 | 0.08 | 0.20 | 41–87 | 0.05 | 0.61 | 0.07 | 0.56 |
| CommonsenseQA | 0.75 | 0.07 | 0.17 | 37–85 | 0.11 | 0.57 | 0.02 | 0.54 |
| SIQA | 0.73 | 0.05 | 0.17 | 24–53 | 0.09 | 0.23 | 0.03 | 0.40 |
| *Knowledge & Understanding* | | | | | | | | |
| IFEval | 0.75 | 0.11 | 0.23 | 37–87 | 0.10 | 0.63 | 0.10 | 0.58 |
| EQ-Bench | 0.75 | 0.08 | 0.27 | 17–82 | 0.17 | 0.53 | 0.07 | 0.56 |
| IFBench | 0.71 | 0.11 | 0.31 | 11–32 | 0.08 | 0.51 | 0.07 | 0.55 |
| SuperGPQA | 0.79 | 0.05 | 0.34 | 9–40 | 0.08 | 0.43 | 0.07 | 0.55 |
| MMLU-Pro | 0.65 | 0.10 | 0.40 | 27–71 | 0.17 | 0.36 | 0.18 | 0.51 |

Table 1: Comprehensive benchmark quality metrics across three domains. **CBRC**: Cross-Benchmark Ranking Consistency (Kendall's $\tau$ with peer benchmarks). **DS**: Discriminability Score; Range shows model performance spread (min–max %). **CAD**: Capability Alignment Deviation (higher is better). **BQS**: Combined Benchmark Quality Score. $\sigma$ denotes standard deviation.

push items toward the difficulty frontier where inversions become more likely; and (2) benchmarks with objective, machine-checkable evaluation criteria consistently achieve higher CAD scores than those relying on more subjective scoring.

### 5.2 Model Performance Analysis

Table 2 reveals clear within-family capability hierarchies across all four model families, validating our within-family CAD computation approach. The selective evaluation results further show that our quality-based instance selection preserves the relative ordering of models with an average Kendall's $\tau$ of 0.93 while using only 35% of the original instances; larger models within each family consistently show positive $\Delta$ values on selective benchmarks, indicating that high-quality instances better differentiate capable models, with rank shifts confined to mid-tier models.

### 5.3 Held-Out Model Validation

To validate that our metrics generalize beyond the models used in their computation, we evaluate on held-out models that were not included in computing CAD, DS, or CBRC metrics. Specifically, we use Qwen2.5-Base (1.5B, 7B, 32B), which are base models without instruction tuning. Table 3 presents these results.

The held-out validation demonstrates strong generalization of our selective benchmark approach. Mathematics shows perfect rank preservation (Avg $|\Delta\text{Rk}| = 0.0$), indicating that selective evaluation ranks held-out models identically to full benchmarks in this domain. General Reasoning and Average scores show moderate variation (1.0), while Knowledge & Understanding shows slightly higher variation (1.3). Notably, the extreme-performing 1.5B model maintains its 14th rank consistently across all domains, demonstrating that our selection method is particularly reliable at the capability distribution tails.

### 5.4 Selective Benchmark Construction

We select test instances with high CAD scores (indicating low inversion rates) and high discriminability contributions, creating filtered benchmarks containing approximately 35% of the original instances.

#### 5.4.1 Selection Ratio Analysis

Figure 2 illustrates the effect of selection ratio on benchmark quality metrics. As the selection ratio increases from 10% to 100%, we observe distinct patterns across the three metrics.

Ranking consistency ($\tau$) increases rapidly from 0.88 at 10% to 0.93 at 35%, then plateaus near 0.99

| Model | Mathematics | | | | General | | | | K&U | | | | Avg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F | S | Δ | Rk | F | S | Δ | Rk | F | S | Δ | Rk | F | S | Δ | Rk |
| *DeepSeek-R1-Distill-Qwen* | | | | | | | | | | | | | | | | |
| 32B | 57.5 | 80.1 | +22.6 | 3→3 | 78.8 | 85.5 | +6.7 | 3→3 | 52.8 | 67.5 | +14.7 | 5→4 | 63.1 | 77.7 | +14.6 | 4→3 |
| 7B | 42.8 | 57.7 | +14.9 | 5→5 | 63.9 | 47.7 | -16.2 | 10→10 | 36.6 | 22.1 | -14.5 | 10→11 | 47.8 | 42.5 | -5.3 | 8→9 |
| 1.5B | 26.9 | 26.0 | -0.9 | 10→10 | 38.3 | 13.5 | -24.8 | 13→13 | 20.4 | 6.1 | -14.3 | 13→13 | 28.5 | 15.2 | -13.3 | 13→13 |
| *Llama-3.1-Instruct* | | | | | | | | | | | | | | | | |
| 70B | 30.1 | 32.1 | +2.0 | 8→9 | 79.8 | 86.5 | +6.7 | 2→2 | 60.9 | 79.8 | +18.9 | 3→3 | 56.9 | 66.1 | +9.2 | 5→5 |
| 8B | 13.3 | 4.2 | -9.1 | 13→13 | 67.9 | 56.6 | -11.3 | 8→9 | 47.5 | 54.2 | +6.7 | 7→8 | 42.9 | 38.3 | -4.6 | 10→10 |
| *Qwen2.5-Instruct* | | | | | | | | | | | | | | | | |
| 72B | 49.4 | 79.3 | +29.9 | 4→4 | 78.7 | 85.0 | +6.3 | 4→4 | 62.0 | 81.3 | +19.3 | 1→2 | 63.4 | 81.9 | +18.5 | 3→2 |
| 7B | 29.5 | 37.7 | +8.2 | 9→8 | 73.3 | 69.4 | -3.9 | 6→6 | 51.4 | 58.3 | +6.9 | 6→6 | 51.4 | 55.1 | +3.7 | 6→6 |
| 1.5B | 13.7 | 6.2 | -7.5 | 12→12 | 54.3 | 30.0 | -24.3 | 11→12 | 29.8 | 20.1 | -9.7 | 11→12 | 32.6 | 18.8 | -13.8 | 11→12 |
| *Qwen3* | | | | | | | | | | | | | | | | |
| 32B | 63.5 | 97.1 | +33.6 | 1→1 | 81.7 | 94.2 | +12.5 | 1→1 | 61.9 | 85.7 | +23.8 | 2→1 | 69.1 | 92.3 | +23.2 | 1→1 |
| 8B | 58.4 | 82.9 | +24.5 | 2→2 | 78.3 | 82.9 | +4.6 | 5→5 | 56.3 | 60.6 | +4.3 | 4→5 | 64.3 | 75.5 | +11.2 | 2→4 |
| 1.7B | 34.9 | 49.7 | +14.8 | 6→6 | 69.6 | 57.5 | -12.1 | 7→8 | 44.8 | 37.2 | -7.6 | 8→9 | 49.8 | 48.1 | -1.7 | 7→8 |
| $\tau$ | 0.96 | | | | 0.96 | | | | 0.85 | | | | 0.93 | | | |
| **Stability** | 0.76 | | | | 0.73 | | | | 0.58 | | | | 0.69 | | | |

Table 2: Model performance comparison between full benchmarks (F) and selective evaluation (S). Δ = score difference, Rk = rank change among all 14 models (Full→Selective). Bottom rows show ranking consistency (Kendall's $\tau$) and stability score per domain.

| Model | Mathematics | | | | General | | | | K&U | | | | Avg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F | S | Δ | Rk | F | S | Δ | Rk | F | S | Δ | Rk | F | S | Δ | Rk |
| 32B | 34.5 | 45.0 | +10.5 | 7→7 | 67.5 | 61.6 | -5.9 | 9→7 | 39.7 | 56.7 | +17.0 | 9→7 | 47.2 | 54.4 | +7.2 | 9→7 |
| 7B | 15.2 | 6.4 | -8.8 | 11→11 | 52.8 | 42.8 | -10.0 | 12→11 | 23.8 | 30.4 | +6.6 | 12→10 | 30.6 | 26.5 | -4.1 | 12→11 |
| 1.5B | 1.5 | 1.1 | -0.4 | 14→14 | 5.8 | 7.8 | +2.0 | 14→14 | 4.2 | 5.3 | +1.1 | 14→14 | 3.9 | 4.7 | +0.8 | 14→14 |
| **Avg \|ΔRk\|** | 0.0 | | | | 1.0 | | | | 1.3 | | | | 1.0 | | | |

Table 3: Held-out model validation using **Qwen2.5-Base** family (1.5B, 7B, 32B). These base models were **not used** in computing CAD, DS, or CBRC metrics, which were derived exclusively from instruction-tuned models. F = Full benchmark score (%), S = Selective evaluation score (%), Δ = score difference, Rk = rank among all 14 models (Full→Selective). Avg |ΔRk| shows the average absolute rank change for held-out models (lower is better).

for higher ratios. Stability shows an inverse pattern, starting high at low selection ratios (where only the most reliable instances are included) and gradually decreasing as more instances are added. DS decreases steadily as selection ratio increases, as the most discriminative instances are selected first. The optimal point at 35% achieves a balanced trade-off: ranking $\tau$ of 0.93, stability of 0.69, and DS of 0.47, substantially better than the full benchmark's stability of 0.59 while maintaining comparable ranking consistency.

#### 5.4.2 Metric Combination Ablation

Table 4 compares metric combinations for instance selection. CAD-only and DS-only filtering each excel on a single axis (stability or discriminability) but underperform on the other, while our CAD+DS combination strictly dominates both single-metric variants and the full benchmark, attaining strong ranking consistency (0.93), the highest stability (0.69), and discriminability (0.47).

The default CAD threshold of 0.15 is chosen to balance ranking consistency and stability; a sweep over alternative thresholds is reported in App. J.

#### 5.4.3 Baseline Comparison

Table 5 compares our selection against baselines. High-accuracy selection provides good stability (0.68) but lower ranking consistency (0.87). Medium-difficulty achieves comparable discriminability (0.47) but lower stability (0.51). Our approach achieves the best balance of stability (0.69) and discriminability (0.47) while maintaining strong ranking consistency (0.93).

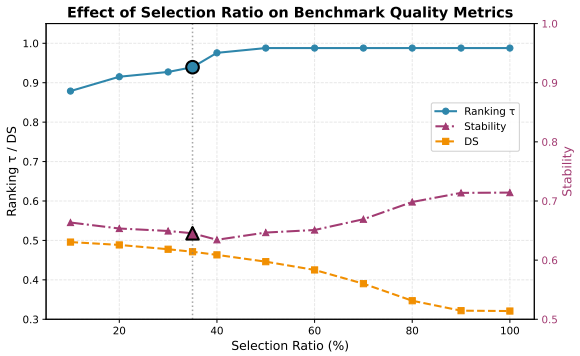


Figure 2: Effect of selection ratio on benchmark quality metrics. The optimal point at 35% (marked) achieves a good balance between ranking consistency ($\tau = 0.93$), stability (0.69), and discriminability (DS = 0.47).

| Method | Level | Rank $\tau$ | Stability | DS |
|---|---|---|---|---|
| *Single Metric* | | | | |
| CAD only (score > 0.15) | Inst | 0.93 | 0.61 | 0.32 |
| DS only (top 35%) | Inst | 0.95 | 0.50 | 0.48 |
| CBRC only (top-2) | Bench | 0.62 | 0.62 | 0.28 |
| *Two Metrics* | | | | |
| **CAD + DS (ours)** | **Inst** | **0.93** | **0.69** | **0.47** |
| CAD + CBRC | Mix | 0.67 | 0.78 | 0.25 |
| DS + CBRC | Mix | 0.59 | 0.58 | 0.33 |
| *Three Metrics* | | | | |
| CAD + DS + CBRC | Mix | 0.65 | 0.76 | 0.29 |
| Full benchmark | – | 1.00 | 0.59 | 0.34 |

Table 4: Metric combination ablation. Level: Inst=Instance, Bench=Benchmark. Our CAD+DS combination (bold) achieves the best balance.

A per-family CAD breakdown (App. K) further confirms that systematic issues such as SIQA's low CAD persist across all four families, while a few benchmarks (e.g., OlympiadBench, EQ-Bench) show family-specific variation worth investigating.

# 6 Discussion and Future Directions

## 6.1 What Makes a High-Quality Benchmark?

Our analysis identifies three key characteristics: (1) *High discriminability*: top benchmarks (AIME, OmniMath, OlympiadBench) achieve DS > 0.7 with wide score ranges; (2) *Strong capability alignment*: benchmarks with CAD > 0.6 respect within-family hierarchies and feature objective evaluation; (3) *Balanced quality profile*: the highest BQS scores emerge from benchmarks balancing multiple dimensions, as exemplified by AIME 2024 (BQS = 0.79).

| Strategy | Rank $\tau$ | Stability | DS |
|---|---|---|---|
| Random selection | $0.91 \pm 0.04$ | $0.59 \pm 0.02$ | $0.34 \pm 0.02$ |
| High accuracy | 0.87 | 0.68 | 0.33 |
| Low accuracy | 0.79 | 0.61 | 0.26 |
| Medium difficulty | 0.93 | 0.51 | 0.47 |
| Longest instances | 0.89 | 0.59 | 0.34 |
| Shortest instances | 0.76 | 0.62 | 0.31 |
| **CAD + DS (ours)** | **0.93** | **0.69** | **0.47** |
| Full benchmark | 1.00 | 0.59 | 0.34 |

Table 5: Comparison of selection strategies. *High/Low accuracy*: instances where most models succeed/fail. *Medium difficulty*: moderate average accuracy. *Longest/Shortest*: most/fewest tokens.

## 6.2 Implications for Benchmark Development

We recommend that benchmark developers: (1) target DS > 0.2 and CAD > 0.6 as minimum thresholds; (2) prefer objective evaluation criteria; (3) consider selective construction using CAD+DS metrics; and (4) monitor family-specific CAD variation as an indicator of potential biases.

## 6.3 Methodological Considerations

**On CBRC and circularity.** Using benchmarks to evaluate benchmarks raises circularity concerns. We mitigate this by selecting widely-adopted reference benchmarks, aggregating across multiple benchmarks, and complementing CBRC with two reference-independent metrics (CAD and DS).

**On model and CAD scope.** Our evaluation spans four model families, with held-out validation confirming generalization. CAD requires multiple model sizes within a family, limiting applicability to single-variant proprietary models, but ensures reliable capability ordering.

# 7 Conclusion

We presented BENCHMETRICS, a framework for evaluating LLM benchmark quality through three complementary metrics: Cross-Benchmark Ranking Consistency, Discriminability Score, and Capability Alignment Deviation. Our evaluation across 15 benchmarks and 11 models reveals significant quality variations among widely-used benchmarks, and demonstrates that selective construction can maintain evaluation fidelity using only 35% of original instances. We hope this framework helps practitioners assess benchmark reliability. Future directions include extending to generation-based evaluations with LLM-as-judge and developing dynamic quality monitoring for benchmark degradation.

## Limitations

Our study has several limitations that suggest directions for future research. First, our evaluation focuses on three domains (mathematics, reasoning, and knowledge understanding); although our metrics are domain-agnostic by design, extending validation to additional domains such as code generation, machine translation, and dialogue systems remains important future work. Second, our analysis is restricted to text-based LLM benchmarks; as multimodal large language models become increasingly prevalent, extending our framework to vision-language, audio-language, and video understanding benchmarks represents a natural next step. Third, while our evaluation spans 11 models across four diverse families, incorporating a broader range of models including proprietary systems would enhance generalizability.

## Ethics Statement

This work involves evaluation of existing public benchmarks and models, and does not introduce new data collection or human subjects research. We use only publicly available benchmarks and evaluate models through their official APIs or publicly released weights. Our framework is intended to improve benchmark quality and thereby contribute to more reliable AI evaluation.

## References


Samuel R Bowman and George E Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

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. *arXiv preprint arXiv:1803.05457*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2368–2378.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4853.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2024. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, and 1 others. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. 2023. vLLM: Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, and 1 others. 2025. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Mathematical Association of America. 2024a. AIME 2024: American invitational mathematics examination. https://www.maa.org/math-competitions/aime-american-invitational-mathematics-examination.

Mathematical Association of America. 2024b. AMC 10/12: American mathematics competitions 2022-2024. https://www.maa.org/math-competitions/amc-1012.

ModelScope Team. 2024. EvalScope: Evaluation framework for large models. https://github.com/modelscope/evalscope.

Samuel J. Paech. 2023. EQ-Bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. 2025. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems*, volume 38.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social IQA: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4463–4473.

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. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Ruijie Xu, Zengzhi Luo, Siyuan Chen, Jing He, Junhe Duan, Fengzhe Wu, Qi Zhang, and Xuanjing Xu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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

## A Experimental Setup Details

We conduct all experiments using the **EvalScope** framework (ModelScope Team, 2024), an open-source evaluation toolkit that provides standardized benchmark implementations and consistent evaluation protocols. For model deployment and inference, we utilize the **vLLM** framework (Kwon et al., 2023), a high-throughput serving engine optimized for large language models.

Table 6 summarizes the key inference parameters used across all experiments. We use the vLLM default sampling configuration (temperature $= 0.7$, top-$p = 0.8$) and set the maximum new tokens to 16384 to accommodate long-form reasoning outputs. To prevent run-to-run sampling variance from contaminating our metrics, each (model, benchmark) pair is evaluated once with a fixed random seed and all of our metrics are aggregated over many instance- and model-level comparisons; bootstrap confidence intervals are reported in App. E. All experiments were conducted on NVIDIA A100 80GB GPUs, with smaller models (1.5B–8B parameters) evaluated using single-GPU deployment and larger models (32B–72B parameters) utilizing multi-GPU tensor parallelism. The complete evaluation across all 15 benchmarks and 14 models required approximately 500 GPU-hours.

| Parameter | Value |
|---|---|
| Temperature | 0.7 |
| Top-p | 0.8 |
| Max new tokens | 16384 |
| Tensor parallelism | Model-dependent |
| GPU memory utilization | 0.90 |

Table 6: Inference configuration for all model evaluations using vLLM framework.

## B CAD Transform Parameter Selection

The Capability Alignment Deviation (CAD) metric applies an exponential transformation to convert raw inversion rates into interpretable scores: $\text{CAD}(B_i) = e^{-\lambda \cdot \text{inv_rate}(B_i)}$. The choice of $\lambda$ affects the sensitivity of the transformed scores to variations in raw inversion rates. We conduct a systematic analysis to select an appropriate value based on five criteria: (1) **Median Mapping**:the median raw inversion rate should map to a score in the range [0.15, 0.35]; (2) **Quality Separation**:different quality levels should exhibit meaningful score differences; (3) **Excellent Quality Reward**:benchmarks with low inversion rates (raw_cad $< 0.03$) should receive high scores ($> 0.65$); (4) **Poor Quality Penalty**:benchmarks with high inversion rates (raw_cad $> 0.25$) should receive low scores ($< 0.10$); and (5) **Dynamic Range**:the transformation should preserve meaningful variation across the main data distribution.

Table 7 presents the evaluation of candidate $\lambda$ values against these criteria. Based on this analysis, we select $\lambda = 12$ as it achieves the highest total score (0.68) by providing strong quality separation (0.93), perfect reward for excellent benchmarks (1.00), full penalty for poor benchmarks (1.00), and complete dynamic range preservation (1.00).

Table 8 shows how raw inversion rates translate to CAD scores with $\lambda = 12$, providing practitioners with concrete reference points for interpreting CAD values in practice.

## C BQS Weight and Normalization Details

The Benchmark Quality Score (BQS) combines three metrics with different native scales. To ensure meaningful aggregation, we apply normalization and empirically-tuned weights.

**CBRC Normalization.** CBRC (Kendall's $\tau$) ranges from $-1$ to 1, while DS and CAD both range from 0 to 1. To align scales, we normalize CBRC using a linear transformation:

$$\widetilde{\text{CBRC}}(B_i) = \frac{\text{CBRC}(B_i) + 1}{2} \tag{9}$$

This maps the CBRC range $[-1, 1]$ to $[0, 1]$, where 0 indicates perfect negative correlation, 0.5 indicates no correlation, and 1 indicates perfect positive correlation.

**Weight Selection.** We assign weights $\alpha = 0.3$, $\beta = 0.3$, and $\gamma = 0.4$ based on the following considerations:

- **CAD receives the highest weight (0.4)** because it operates at the instance level and directly measures whether individual test items respect capability hierarchies, a fundamental property of well-designed benchmarks.
- **CBRC and DS receive equal weights (0.3 each)** as they capture complementary benchmark-level properties: external consistency (CBRC) and internal discriminative power (DS).

**Final Formula.** The complete BQS formula is:

$$\text{BQS}(B_i) = 0.3 \cdot \frac{\text{CBRC}(B_i) + 1}{2} + 0.3 \cdot \text{DS}(B_i) + 0.4 \cdot \text{CAD}(B_i) \tag{10}$$

## D Detailed Model Performance

This section presents the complete performance matrix for all 14 models across the 15 benchmarks, organized by domain.

In the Mathematics domain (Table 9), DeepSeek-R1-Distill-Qwen-32B shows strong performance on competition-style benchmarks, achieving the highest score on AIME 2024 (53.3%). The Qwen3 family demonstrates consistently strong results across all mathematics benchmarks, with Qwen3-32B achieving the highest scores on MATH-500 (87.0%) and AMC 22-24 (67.2%).

| $\lambda$ | Median | Separation | Excellent | Poor | Range | **Total** |
|---|---|---|---|---|---|---|
| 3 | 0.00 | 1.00 | 1.00 | 0.00 | 0.90 | 0.54 |
| 6 | 0.00 | 1.00 | 1.00 | 0.38 | 1.00 | 0.61 |
| 9 | 0.00 | 1.00 | 1.00 | 0.70 | 1.00 | 0.66 |
| 12 | 0.00 | 0.93 | 1.00 | 1.00 | 1.00 | **0.68** |
| 15 | 0.00 | 0.72 | 0.70 | 1.00 | 1.00 | 0.57 |
| 18 | 0.00 | 0.54 | 0.70 | 1.00 | 1.00 | 0.53 |
| 24 | 0.00 | 0.30 | 0.97 | 1.00 | 1.00 | 0.52 |
| 30 | 0.00 | 0.17 | 0.81 | 1.00 | 1.00 | 0.45 |

Table 7: Lambda parameter selection analysis. **Median**: median raw_cad maps to 0.15–0.35 score range; **Separation**: quality level separation; **Excellent**: excellent quality (raw_cad $< 0.03$) receives high score; **Poor**: poor quality (raw_cad $> 0.25$) receives low score; **Range**: dynamic range in main data distribution. The weighted total uses coefficients 0.30, 0.25, 0.20, 0.15, 0.10 respectively.

| **Raw CAD** | **Transformed Score** |
|---|---|
| 0.00 | 1.000 |
| 0.02 | 0.787 |
| 0.03 | 0.698 |
| 0.05 | 0.549 |
| 0.08 | 0.383 |
| 0.10 | 0.301 |
| 0.12 | 0.237 |
| 0.15 | 0.165 |
| 0.20 | 0.091 |
| 0.25 | 0.050 |
| 0.30 | 0.027 |
| 0.40 | 0.008 |

Table 8: Raw CAD to transformed score mapping with $\lambda = 12$. Quality levels: Excellent (raw_cad $< 0.03$), Good (0.03–0.08), Acceptable (0.08–0.15), Concerning (0.15–0.25), Poor ($> 0.25$).

In General Reasoning (Table 10), Qwen3-32B achieves the highest scores on DROP (85.7%) and ARC (95.0%). The results show clear capability hierarchies across model families, with larger models consistently outperforming their smaller counterparts.

For Knowledge & Understanding (Table 11), the larger instruction-tuned models generally achieve higher performance, with IFEval and EQ-Bench showing clearer capability hierarchies across model families.

## E Statistical Reliability Analysis

We compute 95% confidence intervals for all metrics using bootstrap sampling with 1000 iterations. Table 12 presents these intervals across all benchmarks. The results reveal that CBRC estimates show moderate uncertainty (typical CI width of 0.3–0.5), while CAD estimates are notably more stable (typical CI width $< 0.1$). This stability arises because CAD aggregates over many instance-level comparisons, reducing variance. The DS metric shows higher variability, particularly for smaller benchmarks like AIME 2024 (CI: [0.54, 1.19]), reflecting sensitivity to the specific model set evaluated.

## F Cross-Benchmark Correlation Analysis

We compute pairwise Kendall's $\tau$ correlations between benchmarks within each domain to understand the consistency of model rankings across different evaluation instruments.

In Mathematics (Table 13), we observe high correlations among most benchmarks. MATH-500 and AMC 22-24 show strong correlation ($\tau = 0.88$), while OlympiadBench and OmniMath form a nearly perfectly correlated pair ($\tau = 0.99$). AIME 2024 shows moderate correlations with other benchmarks ($\tau \approx 0.62$–$0.71$), reflecting its unique difficulty level.

In General Reasoning (Table 14), DROP and BBH show the highest correlation ($\tau = 0.85$), both requiring complex reasoning. SIQA and CommonsenseQA show strong alignment ($\tau = 0.80$), as both focus on social and commonsense understanding.

The Knowledge & Understanding domain (Table 15) exhibits a relatively uniform correlation

| Model | MATH-500 | AIME 2024 | AMC 22-24 | OlympiadBench | OmniMath |
|---|---|---|---|---|---|
| DeepSeek-R1-Distill-Qwen-1.5B | 68.4 | 16.7 | 35.8 | 6.8 | 6.7 |
| DeepSeek-R1-Distill-Qwen-7B | 73.6 | 16.7 | 39.6 | 49.4 | 34.8 |
| DeepSeek-R1-Distill-Qwen-32B | 74.6 | 53.3 | 41.8 | 61.8 | 56.2 |
| Llama-3.1-Instruct-8B | 49.4 | 0.0 | 16.4 | 0.7 | 0.2 |
| Llama-3.1-Instruct-70B | 68.2 | 23.3 | 26.9 | 18.3 | 13.9 |
| Qwen2.5-Instruct-1.5B | 49.8 | 0.0 | 18.7 | 0.0 | 0.0 |
| Qwen2.5-Instruct-7B | 75.2 | 13.3 | 40.3 | 8.1 | 10.4 |
| Qwen2.5-Instruct-72B | 83.4 | 16.7 | 50.7 | 52.2 | 44.1 |
| Qwen3-1.7B | 72.2 | 10.0 | 40.3 | 28.9 | 22.9 |
| Qwen3-8B | 84.6 | 26.7 | 61.9 | 61.8 | 57.1 |
| Qwen3-32B | 87.0 | 36.7 | 67.2 | 64.8 | 62.0 |

Table 9: Mathematics domain: Model performance (%) on each benchmark.

| Model | DROP | ARC | BBH | SIQA | CommonsenseQA |
|---|---|---|---|---|---|
| DeepSeek-R1-Distill-Qwen-1.5B | 42.0 | 56.7 | 30.8 | 24.9 | 37.3 |
| DeepSeek-R1-Distill-Qwen-7B | 66.2 | 83.9 | 69.4 | 37.6 | 62.4 |
| DeepSeek-R1-Distill-Qwen-32B | 76.6 | 93.2 | 89.6 | 51.4 | 83.1 |
| Llama-3.1-Instruct-8B | 66.8 | 86.3 | 67.5 | 42.3 | 76.7 |
| Llama-3.1-Instruct-70B | 87.9 | 93.9 | 85.3 | 49.6 | 82.5 |
| Qwen2.5-Instruct-1.5B | 45.9 | 78.1 | 39.7 | 41.5 | 66.5 |
| Qwen2.5-Instruct-7B | 70.2 | 90.6 | 71.1 | 52.2 | 82.4 |
| Qwen2.5-Instruct-72B | 73.7 | 94.2 | 87.3 | 52.8 | 85.4 |
| Qwen3-1.7B | 67.6 | 89.0 | 72.8 | 45.9 | 72.7 |
| Qwen3-8B | 82.0 | 94.1 | 83.5 | 48.9 | 83.0 |
| Qwen3-32B | 85.7 | 95.0 | 89.9 | 53.7 | 84.4 |

Table 10: General Reasoning domain: Model performance (%) on each benchmark.

structure, with IFEval and EQ-Bench showing strong alignment ($\tau = 0.80$).

## G Discriminability Threshold Sensitivity

The Discriminability Score in Eq. (2) uses a threshold $\epsilon = 0.02$ to define a "practically significant" pairwise difference. To verify that this choice is not fragile, we recompute DS across $\epsilon \in \{0.01, 0.02, 0.03, 0.05\}$. Table 16 reports the resulting per-benchmark DS values; the absolute scores scale proportionally with $\epsilon$, but the *relative ordering of benchmarks is perfectly preserved*: Kendall's $\tau$ between any pair of $\epsilon$ values is 1.00.

## H BQS Weight Sensitivity

The default BQS weights $(\alpha, \beta, \gamma) = (0.3, 0.3, 0.4)$ reflect a prior that CAD (instance-level) is more informative than benchmark-level CBRC and DS. To verify that our conclusions do not hinge on this specific choice, we recompute BQS under three alternative weight configurations: *equal* $(1/3, 1/3, 1/3)$, *DS-heavy* $(0.3, 0.4, 0.3)$, and *CBRC-heavy* $(0.4, 0.3, 0.3)$. Table 17 reports the resulting per-benchmark BQS values. The benchmark ranking remains highly correlated with the default across all three configurations (Kendall's $\tau \in [0.75, 0.88]$), and the top-ranked benchmark (AIME 2024) is invariant.

## I Statistical Significance of Selective Construction

To substantiate the claim that our CAD+DS selection improves over alternative strategies, we report paired bootstrap significance tests ($K = 1000$ iterations) and Cohen's $d$ effect sizes computed over per-benchmark statistics.

Table 18 contrasts our CAD+DS selection against (i) the full benchmark, (ii) single-metric variants (CAD only, DS only), and (iii) random selection. CAD+DS attains a significantly higher Stability Score than the full benchmark ($\Delta = +0.10$, $d = 1.01$, $p = 0.005$), DS-only ($\Delta = +0.19$, $d = 1.86$, $p = 0.004$), and random selection ($\Delta = +0.10$, $d = 0.94$, $p = 0.018$). CAD-only filtering, in contrast, does not significantly improve stability over the full benchmark ($\Delta = +0.02$, $d = 0.22$, $p = 0.087$), confirming that the joint use of CAD and DS is essential.

Table 19 compares CAD+DS against six heuristic instance-selection baselines on both ranking consistency ($\tau$) and Stability. CAD+DS significantly improves over Low-accuracy ($p_\tau = 0.003$),

| Model | IFEval | IFBench | EQ-Bench | SuperGPQA | MMLU-Pro |
|---|---|---|---|---|---|
| DeepSeek-R1-Distill-Qwen-1.5B | 37.1 | 11.0 | 17.1 | 9.8 | 27.1 |
| DeepSeek-R1-Distill-Qwen-7B | 58.1 | 13.1 | 49.8 | 18.0 | 43.8 |
| DeepSeek-R1-Distill-Qwen-32B | 73.6 | 19.0 | 76.8 | 31.7 | 63.0 |
| Llama-3.1-Instruct-8B | 76.7 | 25.6 | 66.7 | 20.8 | 47.6 |
| Llama-3.1-Instruct-70B | 87.2 | 32.2 | 82.1 | 35.5 | 67.4 |
| Qwen2.5-Instruct-1.5B | 41.1 | 14.5 | 48.2 | 17.3 | 27.8 |
| Qwen2.5-Instruct-7B | 74.2 | 23.6 | 72.9 | 28.9 | 57.3 |
| Qwen2.5-Instruct-72B | 86.6 | 32.7 | 78.4 | 40.5 | 71.9 |
| Qwen3-1.7B | 70.6 | 22.5 | 63.0 | 20.5 | 47.4 |
| Qwen3-8B | 84.5 | 30.6 | 76.6 | 28.0 | 61.7 |
| Qwen3-32B | 86.6 | 32.4 | 80.1 | 39.4 | 71.0 |

Table 11: Knowledge & Understanding domain: Model performance (%) on each benchmark.

| Benchmark | CBRC | | | DS | | | CAD | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | 95% CI | $\sigma$ | Mean | 95% CI | $\sigma$ | Mean | 95% CI | $\sigma$ |
| MATH-500 | 0.76 | [0.55, 0.91] | 0.09 | 0.27 | [0.09, 0.52] | 0.12 | 0.80 | [0.76, 0.83] | 0.02 |
| AIME 2024 | 0.66 | [0.39, 0.88] | 0.13 | 0.84 | [0.54, 1.19] | 0.16 | 0.92 | [0.83, 1.00] | 0.04 |
| AMC 22-24 | 0.72 | [0.51, 0.88] | 0.09 | 0.43 | [0.24, 0.65] | 0.11 | 0.69 | [0.62, 0.75] | 0.03 |
| OlympiadBench | 0.79 | [0.61, 0.91] | 0.08 | 0.82 | [0.51, 1.15] | 0.16 | 0.77 | [0.75, 0.79] | 0.01 |
| OmniMath | 0.78 | [0.61, 0.90] | 0.07 | 0.85 | [0.58, 1.16] | 0.16 | 0.78 | [0.75, 0.80] | 0.01 |
| DROP | 0.68 | [0.39, 0.90] | 0.13 | 0.30 | [0.11, 0.54] | 0.12 | 0.80 | [0.80, 0.81] | 0.00 |
| ARC | 0.73 | [0.51, 0.90] | 0.10 | 0.27 | [0.06, 0.51] | 0.12 | 0.94 | [0.93, 0.95] | 0.00 |
| BBH | 0.70 | [0.43, 0.90] | 0.12 | 0.31 | [0.17, 0.46] | 0.08 | 0.82 | [0.82, 0.83] | 0.00 |
| SIQA | 0.70 | [0.43, 0.90] | 0.12 | 0.29 | [0.10, 0.52] | 0.11 | 0.49 | [0.47, 0.51] | 0.01 |
| CommonsenseQA | 0.68 | [0.42, 0.88] | 0.12 | 0.28 | [0.07, 0.51] | 0.12 | 0.79 | [0.77, 0.81] | 0.01 |
| IFEval | 0.69 | [0.47, 0.85] | 0.09 | 0.31 | [0.19, 0.45] | 0.07 | 0.79 | [0.76, 0.82] | 0.02 |
| IFBench | 0.55 | [0.22, 0.80] | 0.15 | 0.32 | [0.23, 0.39] | 0.04 | 0.74 | [0.69, 0.78] | 0.02 |
| EQ-Bench | 0.70 | [0.54, 0.85] | 0.08 | 0.35 | [0.12, 0.61] | 0.12 | 0.77 | [0.71, 0.83] | 0.03 |
| SuperGPQA | 0.65 | [0.39, 0.83] | 0.11 | 0.39 | [0.22, 0.57] | 0.09 | 0.67 | [0.67, 0.68] | 0.00 |
| MMLU-Pro | 0.59 | [0.21, 0.85] | 0.17 | 0.44 | [0.25, 0.64] | 0.10 | 0.54 | [0.53, 0.54] | 0.00 |

Table 12: Bootstrap 95% confidence intervals for all metrics across benchmarks (1000 iterations).

Shortest ($p_\tau = 0.004$), Random ($p_{\text{Stab}} = 0.013$), and Medium-difficulty ($p_{\text{Stab}} = 0.001$) selection, and is statistically indistinguishable from High-accuracy on stability while still improving $\tau$ ($p_\tau = 0.034$).

## J CAD Threshold Sensitivity for Instance Selection

Table 20 reports the effect of varying the CAD threshold used for instance selection. A very restrictive threshold (0.40) attains perfect stability (1.00) but lowers ranking consistency (0.87) due to insufficient retained instances. The default threshold of 0.15 strikes an optimal balance, achieving strong ranking consistency (0.93) and good stability (0.69) while retaining 84% of the original instances.

## K CAD Breakdown by Model Family

Table 21 reports the CAD score of each benchmark restricted to each individual model family. Llama achieves near-perfect CAD on several benchmarks (AIME 2024: 1.00, OmniMath: 0.98), while Qwen3 shows higher variation across benchmarks (OlympiadBench: 0.39, EQ-Bench: 0.43). SIQA exhibits consistently low CAD across all families (0.20–0.27), indicating inherent design issues rather than a family-specific artifact.

## L Stability Score under Different Bootstrap Budgets

The Stability Score in Eq. (5) is computed with $K = 100$ bootstrap iterations. To check that the choice of $K$ does not materially affect our conclusions, we recompute Stability for $K \in \{50, 100, 200, 500\}$. Table 22 shows that the per-domain Stability values are essentially identical across all four budgets (per-domain standard deviation $\leq 0.005$, average Stability $= 0.69$ for every $K$), confirming that $K = 100$ provides a reliable estimate.

| | MATH-500 | AIME 2024 | AMC 22-24 | OlympiadBench | OmniMath |
|---|---|---|---|---|---|
| MATH-500 | 1.00 | 0.65 | 0.88 | 0.75 | 0.75 |
| AIME 2024 | – | 1.00 | 0.62 | 0.71 | 0.69 |
| AMC 22-24 | – | – | 1.00 | 0.70 | 0.70 |
| OlympiadBench | – | – | – | 1.00 | 0.99 |
| OmniMath | – | – | – | – | 1.00 |

Table 13: Mathematics domain: Pairwise Kendall's $\tau$ correlation between benchmarks.

| | DROP | ARC | BBH | SIQA | CommonsenseQA |
|---|---|---|---|---|---|
| DROP | 1.00 | 0.74 | 0.85 | 0.58 | 0.56 |
| ARC | – | 1.00 | 0.71 | 0.76 | 0.74 |
| BBH | – | – | 1.00 | 0.65 | 0.63 |
| SIQA | – | – | – | 1.00 | 0.80 |
| CommonsenseQA | – | – | – | – | 1.00 |

Table 14: General Reasoning domain: Pairwise Kendall's $\tau$ correlation between benchmarks.

| | IFEval | IFBench | EQ-Bench | SuperGPQA | MMLU-Pro |
|---|---|---|---|---|---|
| IFEval | 1.00 | 0.75 | 0.80 | 0.62 | 0.57 |
| IFBench | – | 1.00 | 0.54 | 0.47 | 0.43 |
| EQ-Bench | – | – | 1.00 | 0.80 | 0.67 |
| SuperGPQA | – | – | – | 1.00 | 0.69 |
| MMLU-Pro | – | – | – | – | 1.00 |

Table 15: Knowledge & Understanding domain: Pairwise Kendall's $\tau$ correlation between benchmarks.

| **Benchmark** | $\epsilon = 0.01$ | $\epsilon = 0.02^*$ | $\epsilon = 0.03$ | $\epsilon = 0.05$ |
|---|---|---|---|---|
| AIME 2024 | 0.85 | 0.74 | 0.66 | 0.55 |
| OmniMath | 0.90 | 0.79 | 0.70 | 0.59 |
| OlympiadBench | 0.87 | 0.76 | 0.67 | 0.56 |
| AMC 22-24 | 0.42 | 0.36 | 0.31 | 0.25 |
| MATH-500 | 0.20 | 0.16 | 0.13 | 0.09 |
| ARC | 0.14 | 0.11 | 0.09 | 0.06 |
| BBH | 0.30 | 0.25 | 0.21 | 0.17 |
| DROP | 0.24 | 0.20 | 0.17 | 0.13 |
| CommonsenseQA | 0.21 | 0.17 | 0.14 | 0.10 |
| SIQA | 0.21 | 0.17 | 0.14 | 0.10 |
| IFEval | 0.28 | 0.23 | 0.20 | 0.15 |
| EQ-Bench | 0.32 | 0.27 | 0.23 | 0.18 |
| IFBench | 0.37 | 0.31 | 0.27 | 0.21 |
| SuperGPQA | 0.40 | 0.34 | 0.30 | 0.24 |
| MMLU-Pro | 0.47 | 0.40 | 0.35 | 0.28 |
| **Rank $\tau$ vs $\epsilon = 0.02$** | **1.00** | **1.00** | **1.00** | **1.00** |

Table 16: Discriminability Score under different $\epsilon$ thresholds. $^*$ denotes the default. The benchmark ranking induced by DS is perfectly stable across all thresholds.

| Benchmark | Default (.3,.3,.4) | Equal ($\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$) | DS-heavy (.3,.4,.3) | CBRC-heavy (.4,.3,.3) |
|---|---|---|---|---|
| AIME 2024 | 0.79 | 0.78 | 0.78 | 0.78 |
| OmniMath | 0.74 | 0.76 | 0.76 | 0.77 |
| OlympiadBench | 0.73 | 0.75 | 0.75 | 0.76 |
| AMC 22-24 | 0.55 | 0.56 | 0.54 | 0.59 |
| MATH-500 | 0.55 | 0.54 | 0.51 | 0.57 |
| ARC | 0.65 | 0.62 | 0.57 | 0.65 |
| BBH | 0.60 | 0.59 | 0.56 | 0.62 |
| DROP | 0.56 | 0.55 | 0.52 | 0.58 |
| CommonsenseQA | 0.54 | 0.54 | 0.50 | 0.57 |
| SIQA | 0.40 | 0.42 | 0.40 | 0.47 |
| IFEval | 0.58 | 0.58 | 0.54 | 0.61 |
| EQ-Bench | 0.56 | 0.56 | 0.53 | 0.59 |
| IFBench | 0.55 | 0.56 | 0.53 | 0.59 |
| SuperGPQA | 0.54 | 0.55 | 0.53 | 0.59 |
| MMLU-Pro | 0.51 | 0.53 | 0.52 | 0.56 |
| **Rank $\tau$ vs default** | **1.00** | **0.88** | **0.75** | **0.85** |

Table 17: BQS sensitivity to the three weight parameters. The benchmark ranking induced by BQS remains highly correlated with the default across alternative weight configurations, and the top-ranked benchmark (AIME 2024) is invariant.

| Comparison | Stab. (Ours) | Stab. (Base) | $\Delta$ | Cohen's $d$ | $p$ |
|---|---|---|---|---|---|
| CAD+DS vs. Full | 0.69 | 0.59 | $+0.10$ | 1.01 | 0.005 |
| CAD+DS vs. CAD only | 0.69 | 0.61 | $+0.08$ | 0.91 | 0.023 |
| CAD+DS vs. DS only | 0.69 | 0.50 | $+0.19$ | 1.86 | 0.004 |
| CAD+DS vs. Random | 0.69 | 0.59 | $+0.10$ | 0.94 | 0.018 |
| CAD only vs. Full | 0.61 | 0.59 | $+0.02$ | 0.22 | 0.087 |
| DS only vs. Full | 0.50 | 0.59 | $-0.09$ | 1.04 | 0.010 |

Table 18: Significance of metric-combination effects on Stability Score. $p$-values are from paired bootstrap (1000 iterations); Cohen's $d$ is computed over per-benchmark stability scores.

| Comparison | $\tau$ (Ours) | $\tau$ (Base) | $p(\tau)$ | Stab. (Ours) | Stab. (Base) | $p$(Stab) |
|---|---|---|---|---|---|---|
| CAD+DS vs. Random | 0.93 | 0.91 | 0.255 | 0.69 | 0.59 | 0.013 |
| CAD+DS vs. High accuracy | 0.93 | 0.87 | 0.034 | 0.69 | 0.68 | 0.102 |
| CAD+DS vs. Low accuracy | 0.93 | 0.79 | 0.003 | 0.69 | 0.61 | 0.026 |
| CAD+DS vs. Medium diff. | 0.93 | 0.93 | 0.105 | 0.69 | 0.51 | 0.001 |
| CAD+DS vs. Longest | 0.93 | 0.89 | 0.034 | 0.69 | 0.59 | 0.016 |
| CAD+DS vs. Shortest | 0.93 | 0.76 | 0.004 | 0.69 | 0.62 | 0.042 |

Table 19: Significance of CAD+DS over heuristic selection baselines. $p(\tau)$ and $p$(Stab) are $p$-values for ranking-consistency and Stability differences, respectively, from paired bootstrap.

| Threshold | Retained | $\tau$ | Stab. |
|---|---|---|---|
| 0.40 | 58% | 0.87 | 1.00 |
| 0.15 | 84% | 0.93 | 0.69 |
| 0.05 | 95% | 0.96 | 0.55 |
| 0.01 | 98% | 0.98 | 0.52 |
| None | 100% | 0.95 | 0.50 |

Table 20: CAD threshold sensitivity analysis. Higher threshold is more restrictive (fewer instances retained).

| Benchmark | DeepSeek | Llama | Qwen2.5 | Qwen3 |
|---|---|---|---|---|
| MATH-500 | 0.48 | 0.58 | 0.74 | 0.67 |
| AIME 2024 | 0.67 | 1.00 | 1.00 | 0.88 |
| AMC 22-24 | 0.37 | 0.58 | 0.57 | 0.43 |
| OlympiadBench | 0.54 | 0.93 | 0.91 | 0.39 |
| OmniMath | 0.61 | 0.98 | 0.84 | 0.39 |
| DROP | 0.58 | 0.76 | 0.58 | 0.62 |
| ARC | 0.80 | 0.92 | 0.87 | 0.93 |
| BBH | 0.75 | 0.63 | 0.62 | 0.62 |
| SIQA | 0.24 | 0.21 | 0.20 | 0.27 |
| CommonsenseQA | 0.57 | 0.54 | 0.59 | 0.55 |
| IFEval | 0.49 | 0.72 | 0.75 | 0.66 |
| IFBench | 0.62 | 0.52 | 0.49 | 0.44 |
| EQ-Bench | 0.60 | 0.61 | 0.54 | 0.43 |
| SuperGPQA | 0.54 | 0.39 | 0.35 | 0.44 |
| MMLU-Pro | 0.57 | 0.51 | 0.55 | 0.54 |

Table 21: CAD breakdown by model family (transformed scores, higher is better). Values represent $e^{-\lambda \cdot \text{inv_rate}}$ with $\lambda = 12$, where 1.0 indicates perfect alignment.

| $K$ | **Mathematics** | **General Reasoning** | **Knowledge & Understanding** | **Average** |
|---|---|---|---|---|
| 50 | 0.77 | 0.73 | 0.57 | 0.69 |
| 100 | 0.76 | 0.73 | 0.58 | 0.69 |
| 200 | 0.77 | 0.72 | 0.58 | 0.69 |
| 500 | 0.76 | 0.73 | 0.57 | 0.69 |
| **Std** | **0.005** | **0.004** | **0.005** | **0.000** |

Table 22: Stability Score across different bootstrap iteration counts $K$. Per-domain values are stable, confirming that $K = 100$ is sufficient.