Why Difficulty Alignment Matters for Model Evaluation: Capability Alignment in LLM Benchmarks

Anonymous ACL submission

Abstract

Benchmark quality metrics such as discriminability and saturation are typically reported as stable properties of datasets. We argue they are not: these metrics are computed on specific model populations and vary substantially across them. This population-dependence is rarely acknowledged in benchmark reports or leaderboards, yet it is a fundamental source of variation in how benchmark quality should be interpreted.

We formalize this position as the Capability Alignment Hypothesis: benchmark informativeness depends on the alignment between item difficulty and the capability distribution of evaluated models. Empirically, we show that discriminability follows an inverted-U relationship with difficulty, where items that are too easy or too hard for a given population yield weak discrimination. We introduce the Capability Alignment Score (CAS), combining difficulty alignment and ability-consistent discrimination, as a complementary diagnostic signal alongside existing metrics. Experiments across math and reasoning benchmarks confirm that CAS captures alignment-related structure not fully reflected in current measures.

1 Introduction

Benchmark results are central to progress claims in LLM research (Liang et al., 2023). In practice, benchmark scores are often treated as reliable evidence for comparing models, tracking progress, and selecting deployment candidates (Wu et al., 2023; Javaji et al., 2025). However, recent work has shown that benchmark behavior itself can be unstable or misleading: top models can become compressed on leaderboards (saturation), models can be weakly separated, and rankings can vary across benchmark choices assessing the same phenomenon (Hofmann et al., 2025; Alzahrani et al., 2024).

These findings have prompted several lines of benchmark-quality assessment, producing metrics for separability, ranking consistency, and saturation (Qian et al., 2026; Hofmann et al., 2025). Yet these diagnostics primarily characterize *outcomes* of evaluation; they do not explain *when* a benchmark is likely to be informative for a given model population (Bean et al., 2025).

One subtler issue has received little attention: benchmark quality metrics are inherently *population-sensitive*. Discriminability, saturation, and other quality metrics are computed on a specific set of models, yet they are typically reported as properties of the benchmark itself, without reference to the capability distribution of the evaluated population. This practice obscures an important source of variation: the same benchmark may exhibit high discriminability for one model population and low discriminability for another.

In this paper, we focus on one specific mechanism: **difficulty–capability alignment**. Our central observation is that benchmark informativeness depends on where item difficulty lies relative to model capability. When items are too easy for nearly all models, evaluations saturate and discriminability declines (ceiling effects). When items are too hard for nearly all models, discriminability also declines (floor effects). The most informative region typically appears near intermediate difficulty, where models are partially but not uniformly successful, consistent with classical and NLP-adapted Item Response Theory perspectives on measurement information (Cappelleri et al., 2014; Byrd and Srivastava, 2022). Figure 1 provides a conceptual overview of these regimes.

This leads to our **Capability Alignment Hypothesis**: benchmark items are most informative when their difficulty is aligned with the capability frontier of the evaluated models. Under this view, benchmark quality is not an intrinsic property of a dataset alone; it is a relational property between

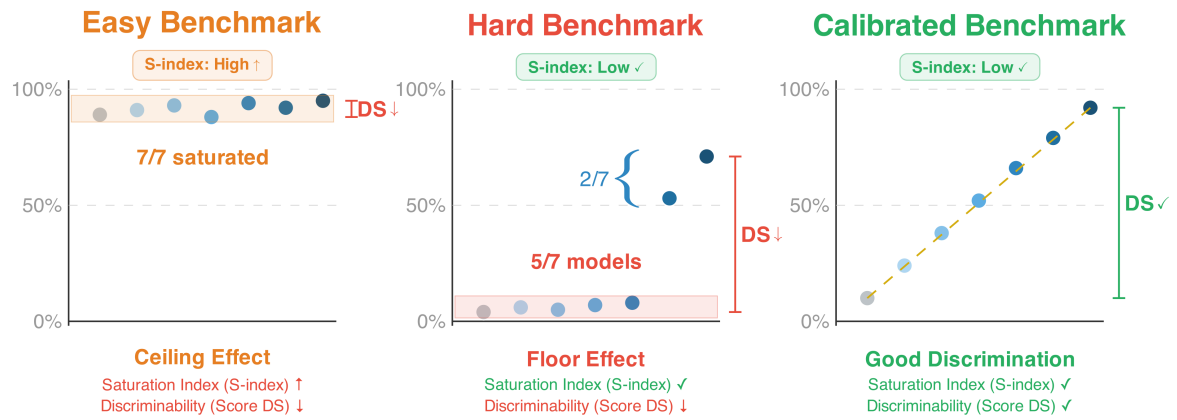

Figure 1: Capability alignment: benchmark informativeness is highest when item difficulty aligns with model capability, and degrades under ceiling or floor regimes.

an item distribution and a model population. The same benchmark can therefore be informative in one capability regime and much less informative in another, a pattern also emphasized in dynamic-benchmark and population-dependent evaluation perspectives (Kiela et al., 2021; Ethayarajh and Jurafsky, 2020). This relational framing implies that benchmark quality assessments should always be reported alongside the capability distribution they were computed on and even considered during benchmark design and items composition, which seems to mostly fall under implicit assumptions from both benchmark creators and users.

To put this into practice, we introduce the **Capability Alignment Score (CAS)**, an item-level alignment signal that combines (i) difficulty positioning (via entropy of item success rate) and (ii) ability-consistent discrimination (via rank correlation with model ability). We aggregate CAS at benchmark level and analyze its relationship with existing quality metrics.

Across math and reasoning benchmarks, we find recurring evidence consistent with this framing: (1) an inverted-U relationship between difficulty and discriminability, (2) strong empirical association between higher CAS, higher DS, and lower saturation (Appendix C), and (3) practical gains from CAS-aware filtering in DS/saturation trade-offs.

Our main contributions are:

- We formalize the Capability Alignment Hypothesis that benchmark quality is a relational property between item difficulty and model capability and provide empirical support via an inverted-U relationship between difficulty and discriminability.
- We propose the Capability Alignment Score (CAS) as a diagnostic and predictive metric to complement existing benchmark-quality metrics.
- We demonstrate that CAS-informed filtering improves discriminability and reduces saturation.

Section 2 reviews related work and the benchmark-quality metrics used throughout. Section 3 presents empirical difficulty-dependent behavior. Section 4 introduces the Capability Alignment Hypothesis and CAS. Sections 5 and 6 describe experiments and results, followed by discussion, limitations, and future directions.

2 Related Work and Adopted Benchmark Metrics

From static leaderboards to broad evaluation ecosystems. NLP benchmarking has evolved from relatively focused static suites to broader, more heterogeneous evaluation pipelines. Foundational benchmarks such as GLUE and SuperGLUE established standardized multi-task evaluation protocols for language understanding (Wang et al., 2018, 2019). More recent efforts expanded scope in both task diversity and capability coverage, including MMLU, BIG-bench, and HELM (Hendrycks et al., 2020; Srivastava et al., 2022; Liang et al., 2023). In parallel, Dynabench argued that static benchmarks can quickly become stale and proposed dynamic data collection to sustain challenge and reduce overfitting to fixed test sets (Kiela et al., 2021).

Benchmark quality metrics and saturation analysis. A growing line of work evaluates benchmark quality itself rather than only model accuracy. Fluid Benchmarking reframes evaluation as benchmark refinement and assesses quality along four dimensions: efficiency, validity, variance, and saturation, showing that IRT-based latent ability estimation plus dynamic item selection can outperform static evaluation pipelines across these axes (Hofmann et al., 2025). Benchmark^2 introduces a complementary multi-metric framework based on cross-benchmark ranking consistency (CRBC), discriminability score (DS), and capability alignment deviation (CAD), and shows that selective benchmark construction can preserve ranking fidelity with fewer items (Qian et al., 2026). In this paper, we use these metrics as reference signals: CRBC (Rank τ) measures agreement between model rankings across related benchmarks; DS measures how strongly a benchmark separates model performance; and CAD (as implemented in Benchmark^2 via transformed inversion rate) assigns higher values to better capability-order alignment. Complementarily, saturation-focused work defines benchmark plateauing as loss of reliable discriminative power among top models and measures it with an uncertainty-aware saturation index (S-index), which captures top-leaderboard compression under uncertainty-aware comparison (Akhtar et al., 2026). For interpretability, CAS, CAD, and S-index are in the $[0, 1]$ range; CRBC is in $[-1, 1]$; DS is non-negative but theoretically unbounded. Higher CRBC/DS/CAD indicate better benchmark quality signals, while a higher S-index indicates stronger saturation, which is undesirable. These studies produce practical diagnostics for leaderboard behavior, but primarily characterize *outcomes* (e.g., compression, separability, and ranking consistency) rather than explicitly modeling how those outcomes depend on item difficulty relative to model capability.

Validity and interpretability of benchmark claims. Beyond metric behavior, recent work emphasizes the validity of claims drawn from benchmark scores. In particular, construct-validity analyses argue that benchmark evidence should be interpreted through the full chain from phenomenon definition to task operationalization, scoring, and downstream claims (Bean et al., 2025). Earlier critiques of leaderboard practice likewise note that score gains do not necessarily translate to user utility or scientific progress (Ethayarajh and Jurafsky, 2020). This perspective motivates moving from "which benchmark has higher scores" to "under what conditions is benchmark-based measurement decision-useful."

Psychometric perspectives and Item Response Theory. Our work is also related to psychometric test theory, especially Item Response Theory (IRT), where measurement quality depends on the relationship between latent ability and item parameters such as difficulty and discrimination (Rasch, 1960; Lord, 1980; Embretson and Reise, 2000). IRT-inspired NLP studies have used this perspective to build evaluation scales, infer item characteristics, and analyze heterogeneity in benchmark instances (Lalor et al., 2016, 2019; Rodriguez et al., 2022). These efforts support the idea that benchmark informativeness is inherently population-dependent; however, they typically do not provide an alignment-focused benchmark-level diagnostic directly tied to modern LLM evaluation and LLM benchmarks quality assessment.

3 Motivating Observations: How Benchmark Behavior Depends on Difficulty

Existing benchmark quality metrics quantify how well benchmarks differentiate model performance, but do not account for how benchmark behavior depends on the interaction between task difficulty and model capability. Even when a benchmark is designed to measure a valid phenomenon, its ability to provide informative comparisons depends on where its items lie relative to the evaluated models.

In this section, we analyze how benchmark behavior varies with item difficulty, and how this affects discriminability and saturation. We observe recurring patterns across multiple benchmarks in our setup, though with some variability due to the limited number of evaluated models.

3.1 Metrics Capture Outcomes, Not Causes

Existing metrics such as DS, S-index, CRBC, and CAD (defined in Section 2) quantify different aspects of benchmark quality. DS measures how well a benchmark separates models in terms of performance, while the S-index captures the degree of compression among top-performing models.

While these metrics describe observable properties of benchmark outcomes, they do not explain when or why these properties arise.

Difficulty-dependent behavior. To investigate this, we analyze benchmark performance as a function of item difficulty. We define the difficulty of an item i as $1 - p_i$, where p_i is the mean accuracy across evaluated models.

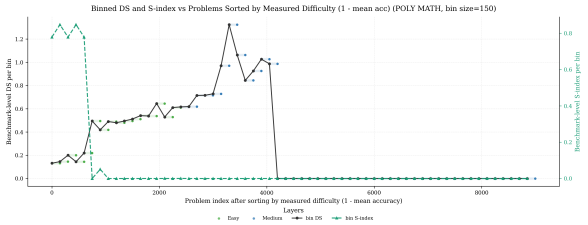

Figure 2: Discriminability Score (DS, left axis, black) and Saturation Index (S-index, right axis, green) across difficulty bins for PolyMath. Item difficulty is defined as $1 - p_i$, where p_i is mean accuracy across models.

Figure 2 shows how DS and S-index vary across difficulty bins for a representative benchmark (an additional GPQA-Diamond example is provided in Appendix A, Figure 7). We observe a recurring pattern:

- For low-difficulty items, most models succeed, leading to low discriminability and higher saturation (ceiling effects).
- As difficulty increases, discriminability tends to improve while saturation decreases.
- For high-difficulty items, most models fail, and discriminability decreases again (floor effects).

These observations are consistent with an inverted-U relationship between difficulty and discriminability, where benchmarks tend to be most informative at intermediate difficulty levels.

Benchmark-defined vs. empirical difficulty. We further analyze benchmark behavior using difficulty labels provided by the benchmark itself, rather than difficulty inferred from model accuracy.

Using benchmark-provided labels yields a similar high-level trend but a different discriminability profile than the model-based definition (Appendix A, Figure 8). In particular, items labeled as "easy" by the benchmark can still exhibit relatively high discriminability when evaluated on the models in our setup.

This discrepancy suggests that difficulty is not an absolute property of benchmark items, but depends on the capability of the evaluated models. As a result, the same item may function as trivial, informative, or overly difficult depending on the model population.

This observation further supports the view that benchmark effectiveness depends on the alignment between item difficulty and model capability.

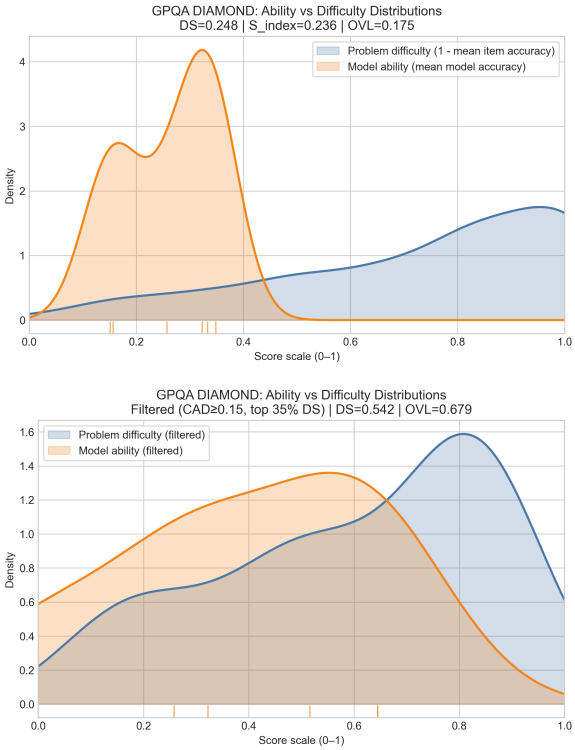

Figure 3: Distributions of model ability (θ_m = mean accuracy) and item difficulty ($1 - p_i$) for GPQA-Diamond, before (top) and after (bottom) filtering based on DS and CAD. OVL denotes the Overlapping Coefficient between the two distributions.

Ability-difficulty interaction. To further examine this interaction, we analyze the distributions of model ability and item difficulty.

Figure 3 and the PolyMath counterpart in Appendix B (Figure 9) show that when the distribution of item difficulty is misaligned with model ability (e.g., mostly too easy or too difficult), benchmarks tend to exhibit lower discriminability. In contrast, when difficulty overlaps with the ability distribution, discriminability tends to increase. We also observe that the Overlapping Coefficient (OVL), which measures the shared area between the two distributions, usually increases after quality-based filtering. However, a few benchmarks show the opposite behavior, where overlap decreases further.

These observations point to the same conclusion: benchmark effectiveness depends on the alignment

between item difficulty and model capability, not difficulty alone.

4 Capability Alignment

The empirical patterns above suggest that benchmark quality is not determined solely by task design or metric choice, but also by how item difficulty interacts with the capability of the evaluated models.

These patterns motivate the Capability Alignment Hypothesis:

Capability Alignment Hypothesis. Benchmark quality is best understood as a relational property: items are most informative when their difficulty aligns with the capability frontier of the evaluated models. Under this view, a benchmark that discriminates well for one model population may saturate or under-discriminate for another.

This hypothesis provides a potential explanation for the observed inverted-U pattern: items that are too easy lead to ceiling effects, while items that are too difficult lead to floor effects. In contrast, items near the capability frontier are more likely to distinguish between models.

A direct implication is that benchmark quality cannot be assessed in isolation from the model population. Benchmark reports that omit this context risk overinterpreting or underinterpreting observed scores. In practice, this means that a reported discriminability score on one model population provides limited evidence about discriminability on a different population.

In the following section, we introduce a metric that allows us to put this principle into practice.

4.1 Capability Alignment Score (CAS)

To measure the Capability Alignment Hypothesis in practice, we introduce the Capability Alignment Score (CAS), a metric that quantifies how well a benchmark aligns with the capability distribution of a given set of models.

Model ability. We first define the ability of a model m on a benchmark as its average accuracy across all items:

$$\theta_m = \frac{1}{N}\sum_{i=1}^{N} y_{i,m}, \tag{1}$$

where $y_{i,m}$ denotes the score of model m on item i. In most benchmarks, $y_{i,m} \in \{0, 1\}$, though the formulation also applies to continuous scores.

Item-level alignment. We define the Capability Alignment Score for an item i as:

$$CAS_i = \underbrace{(-p_i \log_2 p_i - (1-p_i)\log_2(1-p_i))}_{\text{Difficulty Alignment}} \times \underbrace{\max(0,\ \rho(\theta_m, y_{i,m}))}_{\text{Ability-Consistent Discrimination}} \tag{2}$$

where p_i is the mean accuracy of item i, and $\rho(\theta_m, y_{i,m})$ is the Spearman rank correlation between model ability and item outcomes. CAS ranges in $[0, 1]$, where larger values indicate better capability alignment.

The overall CAS of a benchmark is computed as the average across all items:

$$CAS_{\text{benchmark}} = \frac{1}{N}\sum_{i=1}^{N} CAS_i. \tag{3}$$

Interpretation. CAS combines two complementary components: a measure of how well an item is positioned relative to model capability, and a measure of whether the item produces meaningful, ability-consistent differences between models.

Figure 4: Illustration of the two components of CAS. The entropy term is maximized when model responses are split between correct and incorrect outcomes, indicating high informativeness. The correlation term down-weights items where responses are not aligned with model ability (e.g., noisy or inconsistent ordering).

Difficulty alignment (entropy). The first term is the binary entropy of p_i, which captures how informative the item is with respect to the model population. Entropy is low when nearly all models either succeed or fail on the item, and is maximized when the population is split between correct and incorrect responses.

This term therefore favors items that lie near the capability frontier, where some models succeed while others do not, making the item informative for distinguishing between models.

Ability-consistent discrimination (correlation). The second term measures whether item outcomes

are consistent with the overall ability ordering of models. Rather than relying on model size or scale, we define stronger and weaker models based on their empirical performance θ_m, making the formulation model-agnostic.

An item may exhibit a mix of correct and incorrect responses, yet still be uninformative if weaker models perform similarly to stronger ones. Such cases may arise from noise, ambiguity, or random guessing.

To account for this, we compute the rank correlation between model ability and item outcomes. Items with low or negative correlation indicate inconsistent ordering (e.g., inversions), and are downweighted or ignored through the $\max(0, \cdot)$ term. Figure 4 illustrates both components.

5 Experiments

We evaluate the Capability Alignment Hypothesis and the effectiveness of CAS across a diverse set of benchmarks and models. Our experiments are designed to examine (1) how alignment relates to existing benchmark quality metrics, and (2) whether CAS provides a useful signal for identifying informative benchmark items.

Models. We evaluate six models spanning a range of capabilities from approximately 1.5B to 14B parameters. Specifically, we use three models from each of the Qwen3.5 (Qwen Team, 2026) and DeepSeek-Distill-Qwen (DeepSeek-AI, 2025) families:

- **Qwen3.5**: 2B, 4B, 9B
- **DeepSeek-Distill-Qwen**: 1.5B, 7B, 14B

These models provide a range of capabilities while maintaining relative architectural consistency within each family.

Benchmarks. We consider seven benchmarks spanning mathematical reasoning and general knowledge:

- **Mathematics:** AIME 2024, AIME 2025 (Mathematical Association of America, 2025), PolyMath (Wang et al., 2025), MGSM (Shi et al., 2022)
- **General reasoning:** GPQA-Diamond (Rein et al., 2023), MMLU-Redux (Gema et al., 2025), C-Eval (Huang et al., 2023)

These benchmarks cover a range of domains, difficulty levels, and evaluation formats, enabling us to analyze how benchmark behavior varies across different task distributions.

Evaluation setup. We use EvalScope (ModelScope Team, 2024) as the evaluation framework and vLLM (Kwon et al., 2023) for inference. All models are evaluated under a consistent generation setup with a maximum output length of 6144 tokens, temperature 0.7, and top-p 0.8. Reasoning modes are enabled where applicable.

6 Results

This section addresses four questions: (1) what the full, unfiltered benchmarks look like; (2) whether alignment-related patterns emerge across item difficulty; (3) whether CAS adds practical value for benchmark filtering; and (4) an ablation study on the CAS metric.

6.1 Baseline Full-Benchmark Metrics

Table 1 reports benchmark-level metrics before any filtering. CAS is generally low to moderate across datasets, with MGSM as a notable high-CAS case. These baseline values provide the reference point for all subsequent analyses.

Table 1: Benchmark-level quality metrics on full datasets before any filtering.

Dataset	CRBC	DS	CAD	CAS	S-index	Item Count
Math						
AIME24	0.344	1.081	0.587	0.215	0.065	30
AIME25	0.389	0.734	0.301	0.178	0.270	30
MGSM	0.221	0.524	0.365	0.520	0.000	2750
PolyMath	0.408	0.424	0.499	0.203	0.000	9000
Reasoning						
C-Eval	0.400	0.187	0.206	0.263	0.006	1346
GPQA-Diamond	0.000	0.248	0.166	0.187	0.236	198
MMLU-Redux	0.200	0.157	0.548	0.238	0.002	5700

6.2 Does the Alignment Pattern Emerge Across Difficulty?

To examine difficulty-dependent behavior, we bin items by empirical difficulty (defined as $1-p_i$) into four levels (easy, medium, hard, very hard), and compute CAS, DS, and S-index per bin.

Figure 5 shows that DS typically peaks around intermediate difficulty, consistent with the inverted-U pattern observed earlier. CAS generally follows this trend, indicating that informative regions often align with the model capability frontier. In some regions (e.g., the hardest MGSM bin), CAS deviates from DS and becomes more sensitive to saturation-

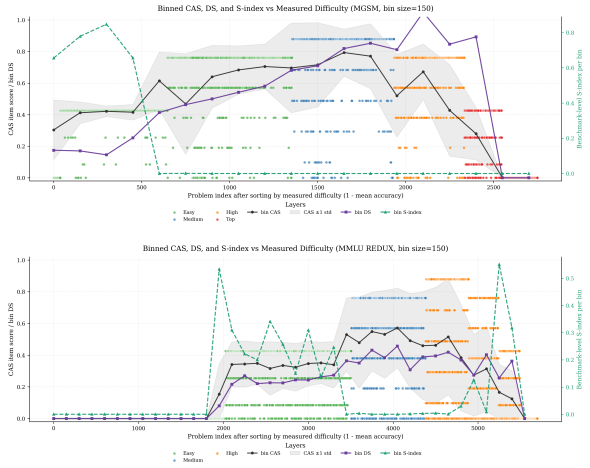

Figure 5: CAS, DS, and S-index across empirical difficulty bins for two representative benchmarks (MGSM, top; MMLU-Redux, bottom).

related effects, suggesting that alignment and discriminability are related but not identical signals.

6.3 Does CAS Add Value for Quality Filtering?

To evaluate practical utility, we follow selective benchmarking from Benchmark^2 (Qian et al., 2026) and compare filtering rules with and without CAS. Our baseline is DS+CAD filtering (Qian et al., 2026), and we then evaluate combinations that add CAS.

We sweep CAS thresholds across all benchmarks and share C-Eval as a representative visualization (Appendix E, Figure 13). Across datasets, the region around CAS > 0.4 provides a practical DS/saturation trade-off, so we use the following thresholds in the filtering experiments:

- DS: keep top 35% items,
- CAD: > 0.15,
- CAS: > 0.4.

Table 2: Average metric changes by filter combination (relative to full-benchmark baselines). Higher is better except for retention and ΔS-index.

Filter	Ret. ↓	ΔDS ↑	ΔCAD ↑	ΔS ↓	Rank τ ↑
DS+CAD+CAS	**20.4%**	+0.162	**+0.582**	**-0.057**	0.782
CAD+CAS	23.1%	+0.151	**+0.582**	**-0.057**	0.807
DS+CAS	33.1%	**+0.219**	-0.012	**-0.057**	**0.964**
DS+CAD	35.4%	+0.131	**+0.582**	+0.005	0.670

Table 2 shows that adding CAS improves the DS/S-index trade-off compared with the DS+CAD baseline, while preserving strong rank consistency. CAS-inclusive filters also retain fewer items, indicating higher compression with equal or better benchmark quality signals. Overall, this suggests that CAS contributes additional information beyond DS and CAD for item selection.

6.4 An Ablation Study

This subsection reports two complementary diagnostics: (1) bootstrap stability of CAS-related effects across benchmarks, and (2) a K-fold validation confirming that the CAS alignment signal is not an artifact of coupling between ability estimates and item-level correlations.

We assess robustness with non-parametric bootstrap over items. For each benchmark, we resample items with replacement (N items per resample), recompute benchmark-level metrics (DS, S-index, CAS), and apply CAS-only filtering (threshold 0.4) to obtain filtered DS. We then report mean and 95% percentile confidence intervals for correlation and improvement estimates.

How CAS–DS correlation is computed. DS is not defined at the item level, so we estimate CAS–DS association at the bin level. In each bootstrap sample, items are grouped into fixed difficulty bins using mean accuracy (p_i), then Spearman correlation is computed across bin-level CAS and DS values.

Benchmark	CAS–DS Corr	ΔDS (Filtered)
AIME24	0.01 [0.00, 0.00]	0.00 [-0.00, 0.00]
AIME25	-0.31 [-0.80, 0.40]	0.14 [0.00, 0.30]
MGSM	-0.02 [-0.10, 0.30]	0.07 [0.05, 0.09]
PolyMath	0.24 [0.10, 0.50]	0.16 [0.13, 0.17]
C-Eval	-0.04 [-0.30, 0.60]	0.33 [0.29, 0.37]
GPQA-Diamond	0.15 [-0.60, 0.80]	0.41 [0.25, 0.60]
MMLU-Redux	0.77 [0.50, 0.90]	0.33 [0.30, 0.34]

Table 3: Bootstrap estimates (mean and 95% CI) for within-benchmark CAS–DS correlation and DS improvement after CAS-based filtering (threshold 0.4).

Table 3 indicates that very small benchmarks (e.g., AIME) yield unstable CAS–DS correlations, with wide intervals and limited gains in some cases. In contrast, medium and large benchmarks—especially reasoning datasets—show more consistent DS improvements after CAS filtering, even when correlation estimates vary.

Assessing statistical leakage in CAS. To address potential coupling between ability estimates and item-level correlations, we recompute CAS using stratified K-fold splits where model ability is estimated on held-out items (details in Appendix D). Figure 6 shows that the alignment pat-

tern remains consistent between raw CAS and K-fold CAS, indicating that the correlation term in CAS is robust to this coupling concern. This generalizes to the other benchmarks as well, where small benchmarks like AIME showed similar variation as GPQA-Diamond, and bigger benchmark showed almost no difference between raw and K-fold CAS.

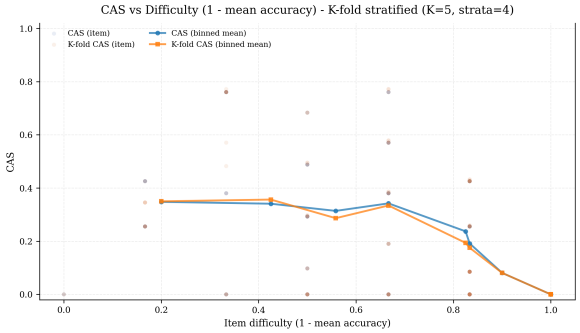

Figure 6: K-fold CAS (orange) and raw CAS (blue) for GPQA-Diamond plotted against empirical item difficulty. Points show individual items; lines show binned means.

7 Discussion

CAS is best interpreted as a diagnostic metric for benchmark structure. It measures difficulty–capability alignment and provides a signal consistently associated with lower saturation and stronger discriminability. At the same time, our ablation analysis (Section 6.4) shows that CAS is a strong candidate to quantify difficulty-capability alignment and how it can provide consistent gains for discriminability as well.

This has practical implications for benchmark design: (1) benchmark reports should include alignment diagnostics (e.g., ability–difficulty distributions), not only aggregate scores; (2) the targeted model population should inform item composition during benchmark construction; and (3) benchmark maintenance should be treated as a lifecycle process, with item pools rebalanced as model frontiers shift. For benchmarks intended to evaluate multiple populations, explicit stratification into capability classes, each with items optimized for within-class discrimination, would clarify where a benchmark is informative and where it saturates.

Finally, this work complements existing meta-evaluation frameworks. Following validity-oriented analyses (Bean et al., 2025), capability alignment is best treated as one validity-relevant dimension: whether items operate in a capability regime where scores carry meaningful discriminative evidence for the target population.

8 Limitations and Future Work

Limitations. This study has several limitations. First, our model pool is relatively small and concentrated in two related families. As a result, the observed alignment patterns may be sensitive to the sampled capability range and may not fully transfer to broader model populations. In particular, benchmarks that appear informative for this population may behave differently for stronger, weaker, or architecturally different models.

Second, several analyses depend on item count and binning granularity. Correlation and trend estimates are less stable on small benchmarks (e.g., AIME-scale datasets), where bootstrap intervals are wide and difficulty-bin statistics can be noisy. These effects limit confidence in fine-grained conclusions for low-sample settings.

Third, CAS is population-dependent by design and currently uses ability estimates derived from the same evaluation pool. While this is consistent with our relational framing, it can introduce dependence between alignment estimates and observed benchmark outcomes. Our current evidence therefore supports strong empirical association, but not strict causal identification.

Future work. Our next step is to scale the study across more model families, broader capability ranges, and additional benchmark domains. A natural extension is to address multi-modal distributions of model abilities and problem difficulty through broader analysis and a more generalized setup, and to develop a framework that automatically discovers capability classes from model-response patterns, then evaluates class-specific item pools for within-class discrimination and staged progression.

A further direction is to evaluate the same benchmark against structurally different model populations, such as comparing how quality and discriminability signals shift across capability tiers, model families, or training paradigms. A related and sharper test is to partition a population by release date relative to the benchmark: models released before the benchmark was published versus those released after. Since post-release models may have been trained with awareness of the benchmark, this split offers a natural probe for contamination effects and can reveal how much of the observed

alignment signal degrades or shifts as the population changes.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, and 1 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *arXiv preprint arXiv:2602.16763*.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, and 1 others. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, and 1 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130.

Joseph C Cappelleri, J Jason Lundy, and Ron D Hays. 2014. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics*, 36(5):648–662.

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

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

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 (EMNLP)*, pages 4846–4853. Association for Computational Linguistics.

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, and 1 others. 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.

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

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. 2025. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in neural information processing systems*, 36:62991–63010.

Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. 2025. Can AI validate science? benchmarking LLMs on claim →Evidence reasoning in AI papers. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2355–2379, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 4110–4124.

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

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657. Association for Computational Linguistics.

John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. 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 4249–4259.

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*.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Mathematical Association of America. 2025. American invitational mathematics examination (AIME) 2024 and 2025 problem sets. Official competition materials.

ModelScope Team. 2024. EvalScope: Evaluation framework for large models. GitHub repository.

Qi Qian, Chengsong Huang, Jingwen Xu, Changze Lv, Muling Wu, Wenhao Liu, Xiaohua Wang, Zhenghua Wang, Zisu Huang, Muzhao Tian, and 1 others. 2026. Benchmark^2: Systematic evaluation of llm benchmarks. *arXiv preprint arXiv:2601.03986*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Y. Pan, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*.

Pedro Rodriguez, Phu Mon Htut, John P Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112.

Freda Shi, Wenxuan Zhou, Yan Xu, Chi Lo, Jackie Cheung, and Xiang Ren. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Ahmad Shoeb, Ali Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, and 1 others. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

A Additional Difficulty Analyses

This section provides additional evidence for the difficulty-dependent behavior described in Sections 3 and 6, showing how the same interpretation transfers across benchmarks and difficulty definitions.

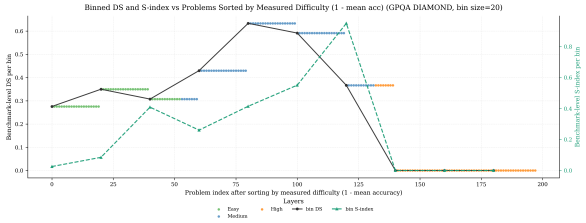

Figure 7: Discriminability Score (DS, left axis, black) and Saturation Index (S-index, right axis, green) across difficulty bins for GPQA-Diamond. Item difficulty is defined as $1 - p_i$, where p_i is mean accuracy across models.

Figure 7 reproduces the inverted-U behavior observed in the main text: low-difficulty regions tend to be saturation-prone, intermediate regions are more discriminative, and very hard regions can lose discriminative power under floor effects. This supports the claim that benchmark informativeness depends on where items sit relative to model capability.

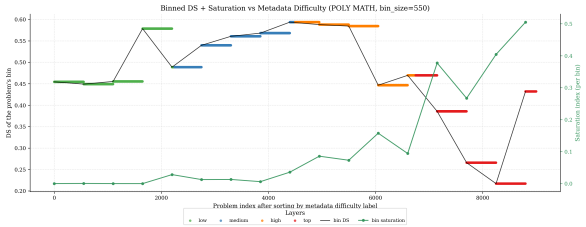

Figure 8: Discriminability Score (DS) and Saturation Index (S-index) across benchmark-provided difficulty levels (PolyMath).

Figure 8 complements this by using benchmark-provided labels instead of empirical difficulty estimated from model accuracy. The qualitative pattern is similar but not identical: items labeled as "easy" can still be discriminative for the model pool used here. This mismatch reinforces the interpretation that item difficulty is population-relative rather than absolute.

B Additional Ability–Difficulty Distributions

This section compares model-ability and item-difficulty distributions before and after filtering for PolyMath, serving as the counterpart to the GPQA-Diamond example in Figure 3.

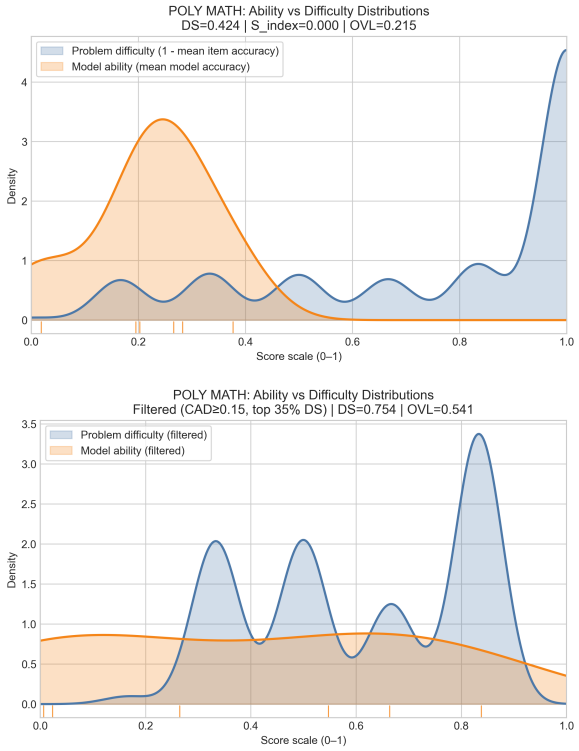

Figure 9: Distributions of model ability (θ_m = mean accuracy) and item difficulty ($1 - p_i$) for PolyMath, before (top) and after (bottom) filtering based on DS and CAD. OVL denotes the Overlapping Coefficient between the two distributions.

Figure 9 shows that quality-aware filtering shifts benchmark mass toward the model ability range, removing regions that are mostly trivial or mostly impossible for the evaluated models. Together with the GPQA-Diamond example in the main text, this supports the view that alignment is about distributional interaction, not difficulty alone.

C Additional Correlation Plots

We test whether CAS is consistent with established quality metrics by measuring the association between CAS, DS, and S-index within each domain. For larger benchmarks, these relationships are computed on benchmark subsets derived from difficulty bins.

C.1 Math Benchmarks

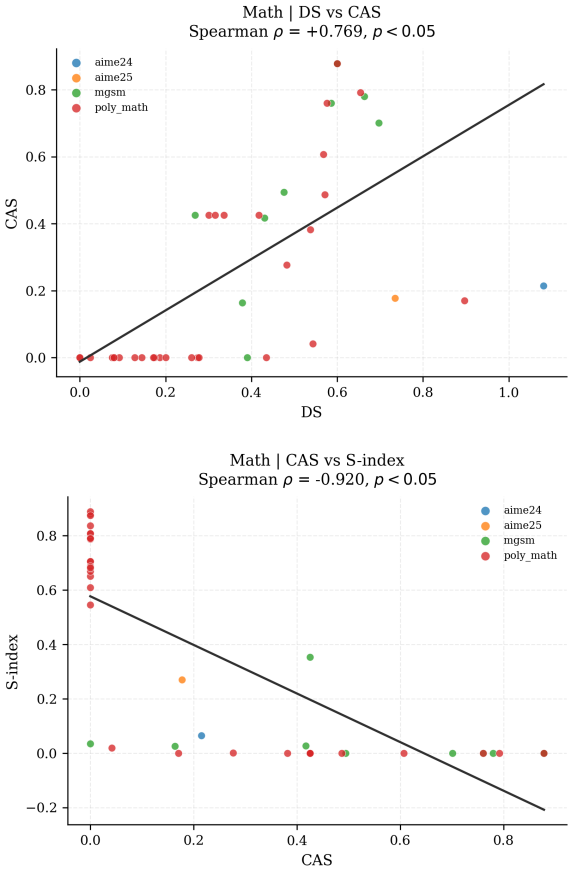

Figure 10: CAS correlation plots for math benchmarks (DS vs. CAS, top; CAS vs. S-index, bottom).

Figure 10 shows a generally positive relationship between CAS and DS, and a generally negative relationship between CAS and S-index, across math benchmarks. Effect strength varies, and estimates are less stable in small-sample settings (see Table 3).

C.2 Reasoning Benchmarks

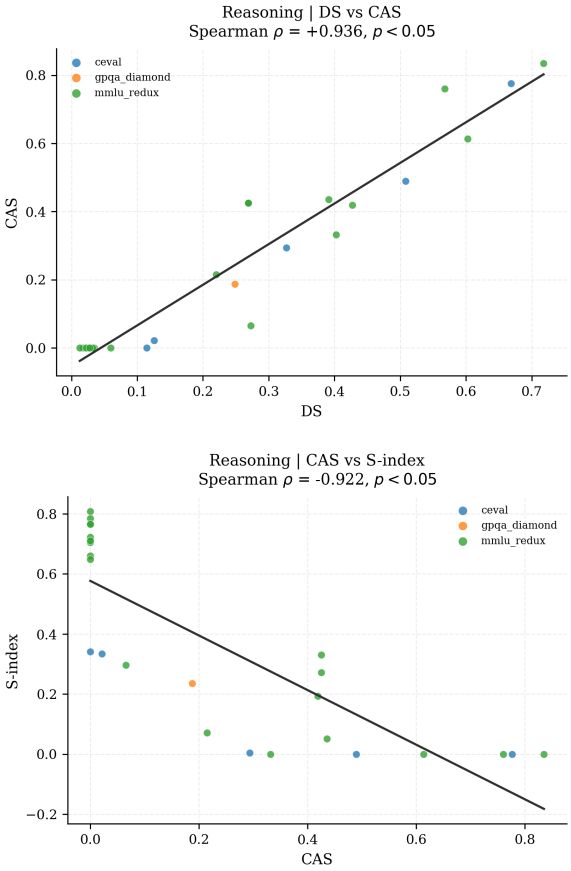

Figure 11: CAS correlation plots for reasoning benchmarks (DS vs. CAS, top; CAS vs. S-index, bottom).

Figure 11 shows the same directional pattern outside the math domain: higher CAS co-occurs with higher DS and lower S-index, matching the sign pattern reported in Section 6. These plots should be read as directional evidence rather than a claim of uniform effect size across all benchmarks.

D K-Fold CAS Methodology

To mitigate potential coupling between ability estimates and item-level correlations, we recompute model ability θ_m using stratified K-fold splits over items. Items are ranked by difficulty $(1 - p_i)$ and partitioned into 4 quantile strata; within each stratum, items are evenly distributed across $K = 5$ folds. For each fold, model ability is estimated on the remaining $K - 1$ folds, and CAS is computed on the held-out items. This produces a per-item K-fold CAS that is decoupled from the item being scored while preserving the difficulty distribution within each fold.

As shown in Figure 6 in the main text, the alignment pattern remains consistent between raw CAS and K-fold CAS, confirming that the correlation term in CAS is robust to this coupling concern.

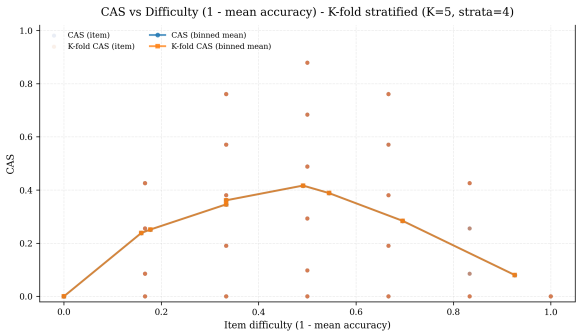

Figure 12: K-fold CAS (orange) and raw CAS (blue) for C-Eval plotted against empirical item difficulty. Points show individual items; lines show binned means.

E Filtering Details

This section documents how the CAS threshold used in Section 6 is selected. We evaluate threshold behavior across all benchmarks and present C-Eval as a representative example of the resulting discriminability/saturation trade-off.

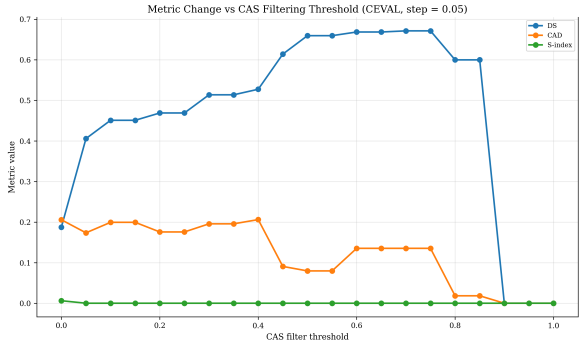

Figure 13: Threshold sweep on C-Eval for CAS-based filtering. CAS > 0.4 yields a favorable balance between higher DS and lower saturation.

Figure 13 shows that CAS > 0.4 provides a practical operating point with improved DS and reduced saturation. This cutoff is consistent with threshold checks across other benchmarks and is used for the filtering comparison in Section 6.