Benchmark Drift: Empirical Characterization of IRT Parameter Stability Across LLM Generations and Architecture Families on Open LLM Leaderboard v2

Keao Xu
Stanford University
keaoxu@stanford.edu

Abstract

LLM benchmark scores implicitly assume *measurement invariance*: the same subtask should separate strong from weak models consistently regardless of when or by whom models were trained. We examine this assumption by introducing the *Beta-2PL* model, a subtask-level extension of 2PL IRT that fits discrimination a_s and difficulty b_s from group mean scores rather than individual binary responses, and applying it to $n = 3{,}326$ models across four architecture families and four quarterly cohorts on Open LLM Leaderboard v2. We find a systematic asymmetry: on BBH (24 subtasks), difficulty rankings are stable across architecture pairs ($\rho_{\text{diff}} \in [0.952, 0.983]$) while discrimination rankings drift ($\rho_{\text{disc}} \in [0.634, 0.933]$, mean dissociation $\Delta\rho = 0.160$); on MATH-HARD (7 subtasks), both parameters are equally stable ($\Delta\rho = 0.024$). We attribute the contrast to construct breadth: cognitively diverse benchmarks are susceptible to discrimination drift as model families improve unevenly across skill domains, while cognitively coherent benchmarks are not. Our findings suggest that benchmark discrimination is architecture- and cohort-sensitive in ways that raw accuracy obscures, with direct implications for anchor-item selection in cross-generation leaderboard comparisons.

1 Introduction

Benchmark scores are treated as measurements of latent model ability, but a valid measurement requires a stable instrument. When the same benchmark subtask separates strong from weak models differently depending on when models were trained or which architecture family they belong to, the instrument is not stable — and cross-generation or cross-architecture score comparisons lose their psychometric grounding.

We introduce a subtask-level extension of the 2PL IRT model — the *Beta-2PL* model — that replaces item-level binary responses with subtask mean scores and fits discrimination a_s and difficulty b_s parameters at the subtask-group level (Section 3). Applying this model to $n = 3{,}326$ models across four architecture families and four quarterly cohorts on Open LLM Leaderboard v2 (Fourrier et al., 2024), we identify a systematic symptom: **discrimination drifts by architecture and cohort while difficulty remains stable**. On Big-Bench Hard (BBH), difficulty rankings are near-perfectly consistent across all architecture pairs ($\rho_{\text{diff}} \in [0.952, 0.983]$) while discrimination rankings diverge substantially ($\rho_{\text{disc}} \in [0.634, 0.933]$). On MATH-HARD, both parameters are equally stable and the gap disappears. The asymmetry reveals that *what a subtask measures* — not just how hard it is — changes as different model families evolve, and does so differently depending on benchmark construct breadth.

Prior IRT work in NLP treats parameters as static (Lalor et al., 2019; Rodriguez et al., 2021; Polo et al., 2024; Atlas et al., 2025). Habba et al. (2025) propose anchor-based cross-generation calibration

Preprint.

but assume drift without characterising its nature. We take a step in this direction by applying the Beta-2PL model and the dissociation statistic $\Delta\rho_{ij}$ (Eq. 4) to examine how discrimination and difficulty rankings shift across cohorts and architecture families — a descriptive analysis we hope motivates more rigorous item-level follow-up.

Contributions. We introduce the Beta-2PL model, a subtask-level IRT extension that fits discrimination and difficulty parameters from group mean scores rather than individual binary responses, making psychometric analysis tractable when only aggregated subtask data are available. Applying it, we document that discrimination is architecture- and cohort-sensitive while difficulty is not — an asymmetry present on cognitively diverse benchmarks (BBH) but absent on cognitively coherent ones (MATH-HARD), with implications for which subtasks can serve as reliable anchors for cross-generation leaderboard comparisons.

All Code is available at https://github.com/ericxu10101/stanford-cs321m-project.

2 Data

Source and benchmarks. We use 2PL IRT parameters estimated from item-level binary responses on Open LLM Leaderboard v2 (Fourrier et al., 2024) across two benchmarks:

- **BBH** (Big-Bench Hard): 24 subtasks covering diverse reasoning skills (logical deduction, causal judgement, object tracking, translation, wordplay, etc.). *Primary benchmark.*
- **MATH-HARD**: 7 subtasks (algebra, geometry, number theory, counting & probability, pre-algebra, pre-calculus, intermediate algebra). *Secondary benchmark.*

Model population and cohorts. We include all $n = 3{,}326$ non-flagged models from the four architecture families of interest, partitioned into four quarterly cohorts by HuggingFace release date (see Appendix A for full breakdown). Cohort sizes grow substantially over time: Q2-2024 ($n = 382$), Q3-2024 ($n = 576$), Q4-2024 ($n = 1{,}005$), Q1-2025 ($n = 1{,}363$). Parameter counts span a wide range: 24% of models have fewer than 7B parameters, 72% fall in the 7–70B range, and 4% exceed 70B. The majority use `bfloat16` precision ($n = 2{,}521$); three mixture-of-experts models are present but not treated separately.

Architecture families. Models are grouped by HuggingFace base class into four families: `LlamaForCausalLM` ($n = 1{,}488$), `Qwen2ForCausalLM` ($n = 1{,}144$), `MistralForCausalLM` ($n = 457$), and `Gemma2ForCausalLM` ($n = 237$). Family sizes are unequal, and Gemma2 in particular has sparse early-cohort coverage ($n = 7$ in Q2-2024), which we flag as a limitation when interpreting Gemma2-specific results.

3 Methods

3.1 Beta-2PL Model

Standard 2PL IRT models the probability that model i answers item j correctly as $P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \sigma(a_j(\theta_i - b_j))$, where θ_i is latent ability, b_j difficulty, and $a_j > 0$ discrimination. Our data is aggregated to the subtask level: for model i and subtask s, the observed response is the mean score $\bar{X}_{is} = \frac{1}{n_s} \sum_{j \in s} X_{ij} \in [0, 1]$. We replace the Bernoulli likelihood with a Beta likelihood suited to bounded continuous responses, retaining the 2PL mean function:

$$\mu_{is} = \sigma(a_s(\theta_i - b_s)), \qquad \bar{X}_{is} \sim \mathrm{Beta}(\mu_{is}\phi,\ (1 - \mu_{is})\phi), \tag{1}$$

where $\phi > 0$ is a global precision parameter. Parameters (a_s, b_s) are defined at the subtask level; interpretations are identical to standard 2PL. Boundary cases are handled by clipping to $[\epsilon, 1 - \epsilon]$ with $\epsilon = 10^{-4}$. The model is fitted separately per (cohort, architecture) group g, yielding $\hat{a}_s^{(g)}$ and $\hat{b}_s^{(g)}$, via a custom extension of `BetaTwoPL` from `torch_measure` (torch_measure, 2025).

3.2 Rank-Order Analysis and Stability Metrics

Parameters from independent fits are identified only up to a linear transformation, so raw values are not comparable across groups. We instead compare *rank orderings* of subtasks within each group,

which are invariant to linear rescaling. For each pair of groups (g_1, g_2) we compute the Spearman rank correlation

$$\rho(g_1, g_2) = r_s\Big(r^{(g_1)},\, r^{(g_2)}\Big)\,, \tag{2}$$

separately for discrimination and difficulty rankings. Per-subtask instability is quantified by rank variance across groups: $\text{RankVar}(s) = \text{SD}(r_s^{(g_1)}, \ldots, r_s^{(g_4)})$. We cross-classify subtasks on cohort and architecture rank variance (partitioned at their medians) into four types: **Invariant** (low on both, anchor candidates), **Temporally unstable**, **Arch-sensitive**, and **Fully unstable**. With $S = 3$ subtasks Spearman ρ is restricted to $\{-1, -0.5, 0, 0.5, 1\}$ regardless of data content; we establish $S \geq 7$ as a practical minimum for informative rank-order analysis.

3.3 Dissociation of Difficulty and Discrimination

To quantify cross-architecture agreement separately for each parameter, we define cohort-averaged vectors $\mathbf{b}^{(i)} = \frac{1}{T}\sum_t \mathbf{b}_t^{(i)}$ and $\mathbf{a}^{(i)} = \frac{1}{T}\sum_t \mathbf{a}_t^{(i)}$, and compute:

$$\rho_{\text{diff}}(i,j) = r_s\Big(\mathbf{b}^{(i)}, \mathbf{b}^{(j)}\Big)\,, \qquad \rho_{\text{disc}}(i,j) = r_s\Big(\mathbf{a}^{(i)}, \mathbf{a}^{(j)}\Big)\,. \tag{3}$$

The *dissociation* for pair (i, j) is:

$$\Delta\rho_{ij} = \rho_{\text{diff}}(i,j) - \rho_{\text{disc}}(i,j), \tag{4}$$

with $\Delta\rho_{ij} > 0$ indicating that difficulty rankings are more universally shared than discrimination rankings. Each architecture's scalar score aggregates over all pairs:

$$\bar{\Delta\rho}_i = \frac{1}{N_{\text{arch}} - 1}\sum_{j \neq i} \Delta\rho_{ij}. \tag{5}$$

We compute Eq. (4) for all $\binom{4}{2} = 6$ pairs on both benchmarks.

4 Results

4.1 Overview: Spearman ρ Across Benchmarks

Figure 1 shows the 4×4 Spearman ρ matrices for difficulty and discrimination rankings across cohorts and architecture families, for BBH (top row) and MATH-HARD (bottom row).

BBH. Difficulty rankings are near-perfectly stable: all cohort pairs $\rho \geq 0.968$; long-range $\rho(\text{Q2}, \text{Q1}) = 0.970$. Architecture pairs $\rho \geq 0.952$ for difficulty. Discrimination is substantially less stable: long-range $\rho = 0.817$, with Q2-2024 as a clear outlier. In the architecture matrix, Llama/Mistral/Qwen2 cluster tightly ($\rho = 0.886$–0.947) while Gemma2 diverges ($\rho = 0.634$–0.840).

MATH-HARD. Both difficulty and discrimination rankings are stable across cohorts (long-range $\rho = 0.964$ for both) and architecture families ($\rho \geq 0.821$ for discrimination, ≥ 0.893 for difficulty). No architecture family is an outlier. The dissociation that characterises BBH is absent here.

4.2 Architecture Clustering: Gemma2 as a BBH-Specific Outlier

Figure 2 presents hierarchical clustering dendrograms constructed from the distance matrices $1 - \rho$, where ρ is the Spearman rank correlation between architecture rank orderings. This analysis reveals striking differences in how architectures group depending on both the benchmark (BBH vs. MATH-HARD) and the metric (difficulty vs. discrimination).

A clear and interpretable structure emerges specifically for BBH discrimination: Mistral and Qwen2 merge first ($\rho = 0.947$), followed by Llama, while Gemma2 branches off at substantially greater distance ($\rho_{\min} = 0.634$), forming a distinct outgroup. For BBH difficulty, however, all four architectures are nearly equidistant with no clear outlier—the clustering pattern is qualitatively different across the two metrics on the same benchmark.

More strikingly, this phenomenon is entirely absent for MATH-HARD. Both the difficulty- and discrimination-based dendrograms show near-equidistant clustering, where no single architecture

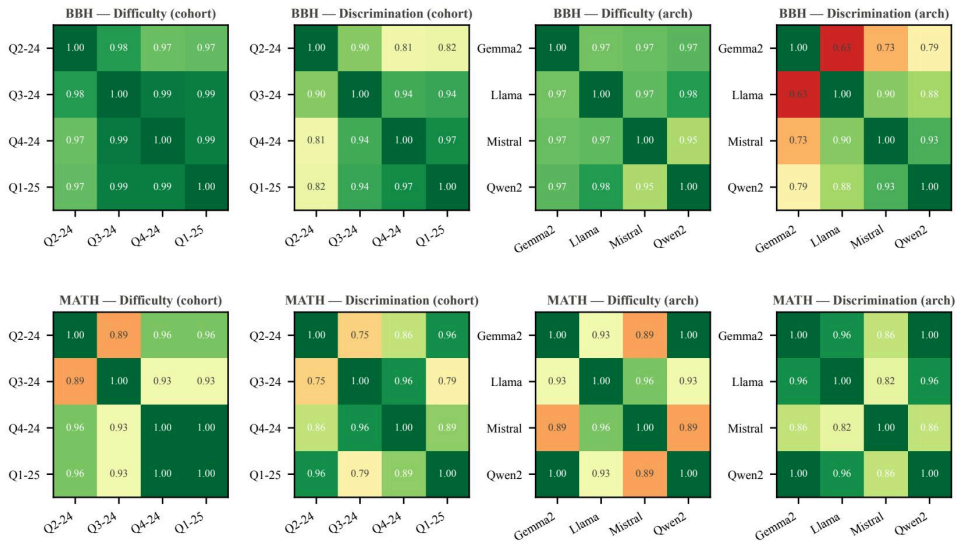

Figure 1: Spearman ρ of subtask rank orderings. **BBH (top row):** difficulty is near-perfect stable across cohorts (≥ 0.968) and architectures (≥ 0.952); discrimination drifts across cohorts (0.814–0.969) and clusters by architecture (0.634–1.000). **MATH-HARD (bottom row):** both parameters are generally stable across all groupings (≥ 0.750 even in the worst case), with no architecture-level outlier.

acts as an outlier. This benchmark-dependent behavior is a noteworthy finding: Gemma2 deviates from other architectures *only* in how its per-subtask discrimination ranks correlate with others on BBH, not in its difficulty ranks, and not on MATH-HARD.

This result suggests that the outlier status of Gemma2 is not an intrinsic property of its architecture (e.g., its parameter count, training regime, or attention mechanism), but rather emerges from a specific interaction between the model and the benchmark's task characteristics. We hypothesize that BBH's diverse collection of reasoning-heavy subtasks may induce different discrimination behaviors in Gemma2 compared to other architectures, a possibility we explore in greater depth in the next section.

We will return to this finding in Section 5 to examine potential explanations, including the role of instruction tuning, tokenization, and emergent reasoning capabilities.

4.3 Temporal Drift Trajectories

Figure 4 (Appendix B) presents discrimination rank trajectories across the four quarterly cohorts, broken out by architecture family (rows) and benchmark (columns). Each panel's highlighted subtasks are identified *per architecture*, so instability that is idiosyncratic to one model family is not obscured by cross-arch averaging.

BBH — cross-architecture consensus. `temporal_sequences` is the most reliable anchor in BBH: it holds rank 1 across all four cohorts for every architecture without exception (variance $= 0$ in all four panels). `causal_judgement` is similarly stable for Llama and Gemma2. At the other extreme, `object_counting` is the most volatile subtask in aggregate (mean rank variance $= 33.9$), ranking among the top-three unstable subtasks for Mistral ($\sigma^2 = 53.6$) and Gemma2 ($\sigma^2 = 58.9$).

BBH — architecture-specific instability. Despite this broad consensus, the *identity* of the next most unstable subtasks diverges sharply by family, revealing an architecture-cohort interaction that cross-arch averages would conceal. For **Llama**, `sports_understanding` is the dominant source

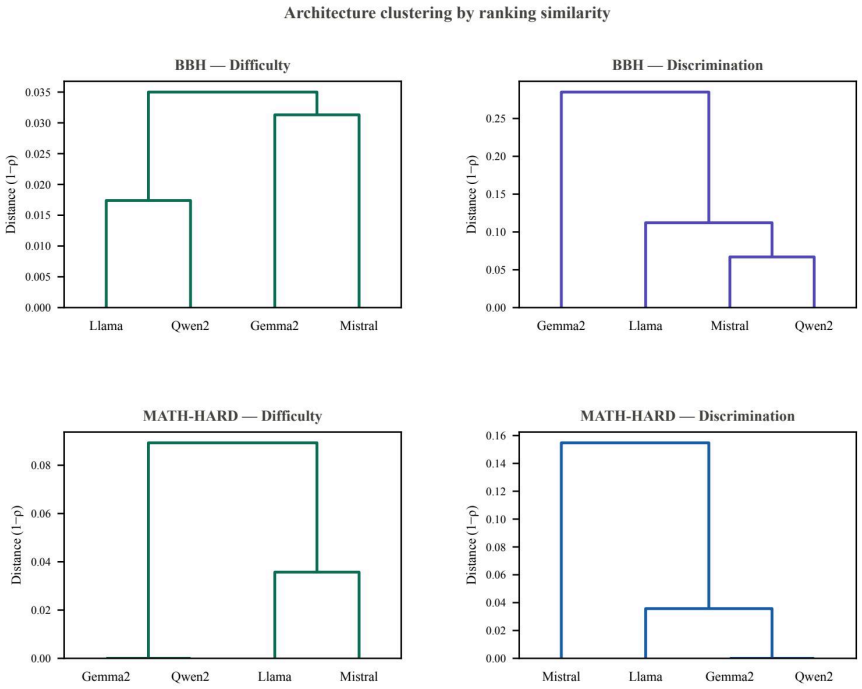

Figure 2: Hierarchical clustering of architectures by ranking similarity (average linkage). **Left column:** BBH. **Right column:** MATH-HARD. **Top row:** difficulty-based rankings. **Bottom row:** discrimination-based rankings. For BBH discrimination (bottom left), Gemma2 branches off as a clear outlier, while all other architectures cluster tightly. This pattern is absent in BBH difficulty and entirely absent in MATH-HARD.

of instability ($\sigma^2 = 8.7$), a subtask that is near-flat for all other families. **Qwen2** exhibits extreme volatility in `geometric_shapes` ($\sigma^2 = 25.0$), which drops from rank 15 in Q2-2024 to rank 5 and stays there, and in `snarks` ($\sigma^2 = 23.0$). **Mistral** and **Gemma2** both show large swings in `hyperbaton` and `object_counting`, but the direction and timing of those swings differ: Mistral's `object_counting` collapses from rank 19 to rank 8 between Q2 and Q4, whereas Gemma2's collapses from rank 24 to rank 6 in the same window before partially recovering. These divergences suggest that temporal drift in discrimination is at least partly driven by architecture-specific learning dynamics rather than benchmark-level difficulty shifts alone.

MATH-HARD — compressed and stable overall. All seven MATH-HARD subtasks show substantially flatter trajectories than BBH across every architecture. The mean cross-arch variance for the most volatile MATH-HARD subtask (`precalculus`, $\bar{\sigma}^2 = 1.75$) is two orders of magnitude below BBH's most volatile subtask, indicating that the MATH-HARD item pool produces discrimination orderings that are largely cohort-invariant. `intermediate_algebra` and `counting_and_prob` are perfectly stable (variance $= 0$) for Llama and Qwen2; `algebra` is perfectly stable for Gemma2. The one notable exception is **Gemma2**'s `precalculus` ($\sigma^2 = 4.9$), which swings from rank 7 in Q2-2024 to rank 2 in Q3-2024 before returning to rank 6 by Q1-2025—the only MATH-HARD trajectory that visually resembles the BBH instability pattern.

Summary. The 4×2 panel decomposition reveals two regimes. BBH discrimination rankings are moderately volatile, with instability concentrated in a small set of subtasks whose identities are partly architecture-specific. MATH-HARD rankings are near-invariant across cohorts for almost all subtask–architecture combinations. Both benchmarks share a single universal anchor: `temporal_sequences` (BBH) achieves zero rank variance across all families, confirming it as an exceptionally consistent discriminator of model capability regardless of training cohort.

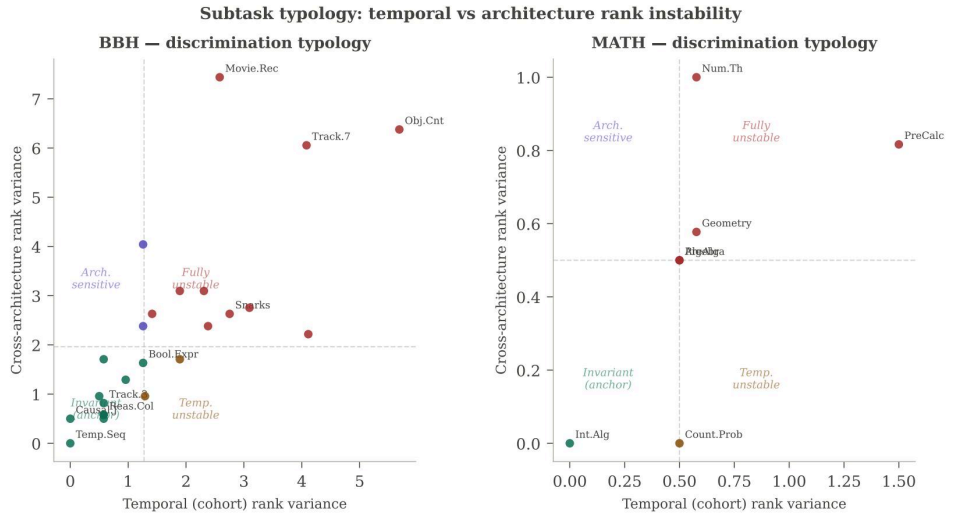

Figure 3: Subtask typology scatter (discrimination). Each dot is one subtask. Quadrants defined by median splits. **BBH:** subtasks spread across all four quadrants; `temporal_sequences` is the sole anchor candidate with zero variance on both axes. **MATH-HARD:** all seven subtasks cluster near the origin; the typology collapses to the invariant quadrant.

4.4 Subtask Typology

Figure 3 plots temporal vs cross-architecture rank variance for BBH (left) and MATH-HARD (right). Table 3 (Appendix C) lists subtask classifications.

For BBH, five subtasks fall in the invariant quadrant and are recommended as anchor candidates. Three subtasks are fully unstable (`object_counting`, `movie_recommendation`, `tracking_7obj`) and should not anchor cross-generation calibration. For MATH-HARD, all seven subtasks cluster in the invariant quadrant — consistent with the near-perfect stability seen in Figures 1 and 4.

4.5 Dissociation of Difficulty and Discrimination Rankings

Full pairwise values for ρ_{diff}, ρ_{disc}, and $\Delta\rho$ across all six architecture pairs on both benchmarks are reported in Table 1 (Equations (3)–(4)).

Table 1: Pairwise dissociation $\Delta\rho_{ij} = \rho_{\text{diff}} - \rho_{\text{disc}}$ for all six architecture pairs on BBH and MATH-HARD. Positive values indicate that difficulty rankings are more cross-architecturally consistent than discrimination rankings. All values use cohort-averaged 2PL estimates (Eq. (**??**)).

		BBH			**MATH-HARD**	
Arch i	Arch j	ρ_{diff}	ρ_{disc}	$\Delta\rho$	ρ_{diff}	ρ_{disc}
Llama	Qwen2	0.983	0.878	+0.104	0.929	0.964
Llama	Mistral	0.970	0.897	+0.073	0.964	0.821
Llama	Gemma2	0.966	0.634	+0.332	0.929	0.964
Qwen2	Mistral	0.952	0.933	+0.019	0.893	0.857
Qwen2	Gemma2	0.971	0.786	+0.185	1.000	1.000
Mistral	Gemma2	0.969	0.725	+0.243	0.893	0.857
Global mean ($\pm$ SD)				0.160 ± 0.106	$\Delta\rho$: 0.024 ± 0.061	

BBH: universal difficulty, idiosyncratic discrimination. Difficulty rankings are highly consistent across all six pairs ($\rho_{\text{diff}} \in [0.952, 0.983]$), confirming that the relative hardness of BBH subtasks

is nearly architecture-invariant. Discrimination rankings are substantially more variable ($\rho_{\text{disc}} \in [0.634,\ 0.933]$), yielding strictly positive dissociation for every pair. The global mean is $\bar{\Delta}\rho = 0.160$ (SD $= 0.106$, $N = 6$ pairs).

Gemma2 is the primary driver of this divergence. Both the Llama–Gemma2 pair ($\Delta\rho = +0.332$, the largest in the table) and the Mistral–Gemma2 pair ($\Delta\rho = +0.243$) stand out, yielding the highest per-architecture score ($\bar{\Delta}\rho_{\text{Gemma2}} = 0.254$, Eq. (5)). Concretely, Llama and Gemma2 rank subtask *difficulties* in near-identical order ($\rho_{\text{diff}} = 0.966$), yet their *discrimination* rankings correlate at only $\rho_{\text{disc}} = 0.634$: a subtask that sharply separates Llama-family models is often uninformative within Gemma2, and vice versa. By contrast, the Qwen2–Mistral pair shows the smallest dissociation ($\Delta\rho = +0.019$, $\rho_{\text{disc}} = 0.933$), consistent with the similar rank-trajectory instability patterns observed for those two families in Section 4.3.

MATH-HARD: dissociation collapses to zero. The MATH-HARD benchmark presents a strikingly different picture. All six pairwise dissociations are near zero ($\Delta\rho \in [-0.036,\ +0.143]$, global mean $= 0.024$, SD $= 0.061$), consistent with the hypothesis $\Delta\rho_{ij} \approx 0\ \forall\, i, j$. The Qwen2–Gemma2 pair achieves $\rho_{\text{diff}} = \rho_{\text{disc}} = 1.00$ ($\Delta\rho = 0.000$): a perfect alignment of both parameter orderings. Two pairs even exhibit negative dissociation (Llama–Qwen2 and Llama–Gemma2, both $\Delta\rho = -0.036$), meaning their discrimination rankings agree *marginally more* than their difficulty rankings — the opposite of the BBH pattern.

Together, these results imply that for MATH-HARD, knowing a subtask's difficulty rank is sufficient to predict its discrimination rank irrespective of architecture: difficulty and discrimination collapse onto a single latent ordering. This is a psychometric signature of a near-unidimensional construct whose item properties generalise across model families, in sharp contrast to BBH, where the two IRT parameters measure structurally distinct and architecture-dependent aspects of subtask quality.

5 Discussion

Why the asymmetry is benchmark-specific. The contrast between BBH and MATH-HARD is not simply a matter of benchmark size (24 vs. 7 subtasks) but of construct breadth. BBH spans cognitively heterogeneous skills—logical deduction, causal reasoning, object tracking, translation, wordplay—that engage distinct capability dimensions. As model families improve across generations, they do so unevenly across these dimensions: a subtask that sharply separated models in Q2-2024 may be near-saturated by Q4-2024, collapsing its discrimination regardless of difficulty. MATH-HARD, by contrast, spans algebra, geometry, number theory, combinatorics, and pre-calculus—domains that, while not identical, share a common substrate of symbolic manipulation and are all saturated at similar rates. Discrimination orderings therefore remain stable not because the subtasks are trivial, but because the underlying capability structure is narrow enough that model families improve along a common frontier. The implication is general: *benchmarks with cognitively diverse subtask pools are structurally more susceptible to discrimination drift than those with coherent, narrow scope*, independent of item count.

The Gemma2 outlier. Gemma2 is the primary driver of BBH dissociation ($\bar{\Delta}\rho_{\text{Gemma2}} = 0.254$, vs. 0.103–0.170 for others) yet behaves indistinguishably from other families on MATH-HARD, ruling out a global architectural explanation. The divergence is directionally structured: Gemma2 finds `movie_recommendation` nearly uninformative ($\hat{a} = 0.602$ vs. 1.776 for others) and `object_counting` weak ($\hat{a} = 0.774$ vs. 1.425), while rating `tracking_7obj` substantially *more* discriminating ($\hat{a} = 1.962$ vs. 0.720). Three candidate explanations — instruction tuning homogenising world-knowledge capabilities across Gemma2 model sizes (Gemma Team, 2024), SentencePiece tokenisation affecting multi-step tracking (Kudo and Richardson, 2018), and ability-range compression deflating discrimination estimates (Hambleton et al., 1991) — are all consistent with the pattern but require item-level data to adjudicate. Gemma2 also exhibits the highest temporal discrimination volatility in BBH (mean subtask $\hat{\sigma} = 0.699$ vs. 0.186–0.527 for others), suggesting the idiosyncrasy is persistent rather than a single-cohort artefact.

Implications for anchor-item calibration. Cross-generation comparability proposals such as Habba et al. 2025 rely on anchor items whose psychometric properties are stable across the conditions being compared. Our results impose two constraints on anchor selection that are not captured by

temporal stability alone. First, an anchor must be invariant on *both* dimensions simultaneously: temporal stability and cross-architecture stability. A subtask stable over time but idiosyncratic to one architecture introduces systematic bias when that architecture is over- or under-represented in a cohort. Only `temporal_sequences` achieves zero rank variance across all four cohorts *and* all four architecture families in BBH; every other subtask drifts on at least one dimension. Second, anchor requirements are benchmark-specific. MATH-HARD's subtask pool approximates a valid anchor set for temporal calibration—all six pairwise dissociations are near zero and rank trajectories are nearly flat—with the single caveat that Gemma2's `precalculus` estimates are anomalously volatile ($\sigma^2_{\text{rank}} = 4.92$) and should be excluded. BBH requires active screening: anchor candidates must be evaluated against both temporal and cross-architecture stability criteria before use.

Limitations. Several limitations constrain the scope of these conclusions. The analysis operates at the subtask level: a subtask classified as stable may contain individual items with substantial drift, and a subtask flagged as unstable may be driven by a small number of outlier items. Item-level 2PL estimation would be needed to distinguish these cases. Four quarterly cohorts provide limited resolution for distinguishing monotone trends from non-monotone fluctuations; longer time series would allow formal trend tests. Discrimination estimates themselves carry estimation uncertainty that propagates into $\Delta\rho$ and rank variance; we report point estimates throughout and do not propagate this uncertainty into confidence intervals for the dissociation statistics. Finally, the cross-benchmark comparison covers four benchmarks from a single leaderboard; whether the BBH–MATH-HARD contrast generalises to other benchmark pairs with similar breadth profiles is an open empirical question.

6 Conclusion

We introduced the Beta-2PL model and applied it to $n = 3{,}326$ models across four architecture families and four quarterly cohorts on Open LLM Leaderboard v2, documenting a systematic difficulty–discrimination dissociation: on BBH, difficulty rankings are stable across architectures ($\rho_{\text{diff}} \in [0.952, 0.983]$) while discrimination rankings drift ($\rho_{\text{disc}} \in [0.634, 0.933]$, $\bar{\Delta}\rho = 0.160$); on MATH-HARD both are equally stable ($\bar{\Delta}\rho = 0.024$). The contrast is explained by construct breadth: diverse benchmarks are susceptible to discrimination drift as model families improve unevenly, while coherent benchmarks are not. Only `temporal_sequences` achieves zero rank variance across all cohorts and all architectures simultaneously, making it the sole reliable BBH anchor candidate; MATH-HARD's pool approximates a valid anchor set with the exception of Gemma2's `precalculus` ($\sigma^2 = 4.92$).

The central implication is that anchor-item calibration for cross-generation comparisons (Habba et al., 2025) requires both temporal and cross-architecture stability — a strictly harder criterion that only a small fraction of BBH subtasks satisfy. Future work should extend the analysis to item level, propagate uncertainty into $\Delta\rho$ estimates, expand to additional benchmarks (e.g., MMLU, GSM8K), and investigate whether Gemma2's idiosyncratic discrimination profile ($\bar{\Delta}\rho_{\text{Gemma2}} = 0.254$) reflects a genuine architectural property or sparse early-cohort coverage ($n = 7$ in Q2-2024).

References

Atlas, S., et al. (2025). Adaptive testing for LLM evaluation. *arXiv:2511.04689*.

Fourrier, C., et al. (2024). Open LLM leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. *arXiv:2408.00118*.

Habba, E., et al. (2025). Growing pains: Extensible LLM benchmarking via fixed parameter calibration. *arXiv:2604.12843*.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP: System Demonstrations*, pp. 66–71.

Lalor, J. P., Wu, H., and Yu, H. (2019). Learning latent parameters without human response patterns. In *Proceedings of EMNLP 2019*.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Polo, F. M., et al. (2024). tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of ICML 2024*.

Rodriguez, P., et al. (2021). Evaluation examples are not equally informative. In *Proceedings of ACL 2021*.

AIMS Foundations (2025). `torch_measure`: A PyTorch-native toolkit for measurement science. https://github.com/aims-foundations/torch_measure.

Truong, S., et al. (2025). Reliable and efficient amortized model-based evaluation. *arXiv:2503.13335*.

Zhou, H., et al. (2025). Lost in benchmarks? Rethinking LLM benchmarking with IRT. In *Proceedings of AAAI 2026*. *arXiv:2505.15055*.

A Model Counts by Cohort and Architecture

Table 2: Model counts per cohort and architecture family.

Cohort	Llama	Qwen2	Mistral	Gemma2	Total
Q2-2024	247	49	79	7	382
Q3-2024	288	83	144	61	576
Q4-2024	415	342	116	132	1,005
Q1-2025	538	670	118	37	1,363
Total	1,488	1,144	457	237	3,326

B Discrimination Rank Trajectories

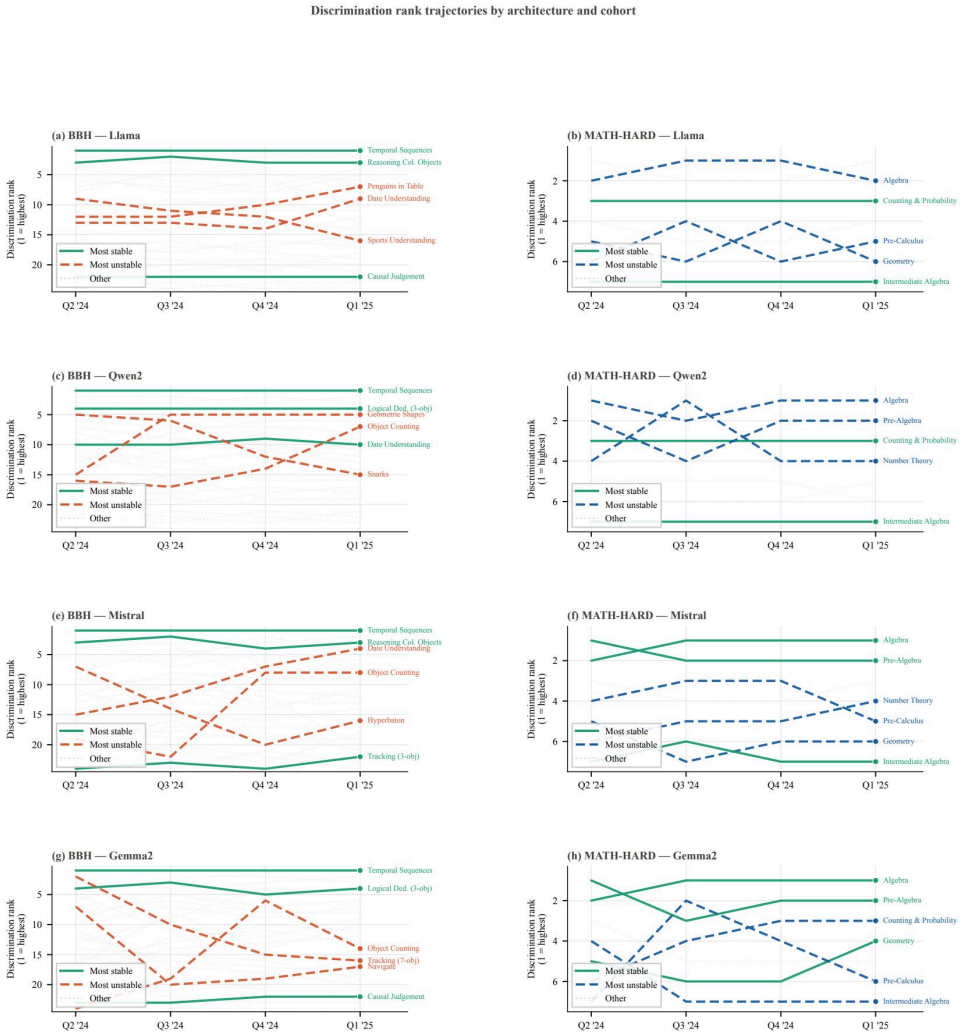

Figure 4: Discrimination rank trajectories across quarterly cohorts, stratified by architecture family (Llama, Qwen2, Mistral, Gemma2; rows) and benchmark (BBH, MATH-HARD; columns). Teal lines = most stable subtasks per architecture; amber lines (BBH) and blue lines (MATH-HARD) = most unstable (dashed, top 3 per panel). Ghost lines = all other subtasks. `temporal_sequences` holds rank 1 in *every* cohort for every architecture (variance $= 0$). `object_counting` is the single most volatile subtask overall (mean cross-arch variance $= 33.9$), but its trajectory differs qualitatively across families (see text).

C BBH Subtask Typology

Table 3: BBH subtask typology (discrimination). Median splits at cohort variance = 1.26 and architecture variance = 1.29.

Invariant	Temporally unstable	Arch-sensitive	Fully unstable
temporal_sequences	object_counting	movie_recommendation	object_counting
causal_judgement	date_understanding	tracking_7obj	movie_recommendation
ruin_names	tracking_7obj	logical_ded_5obj	tracking_7obj
reasoning_col.	hyperbaton	navigate	
tracking_3obj	movie_recommendation		

Scientific Integrity Statement

All empirical claims in this paper are grounded in publicly available data from Open LLM Leaderboard v2 and verified against the raw CSV parameter files included in the code repository. To address plagiarism, every methodological idea is attributed to its original source (Lord, 1980; Birnbaum, 1968; Cronbach, 1972), and the Beta-2PL extension is clearly distinguished from prior work with explicit discussion of what is novel versus adapted. Quantitative claims (Spearman ρ values, rank variances, dissociation statistics) were computed programmatically and cross-checked manually for the most prominent numbers reported in the abstract and results. Potential bias sources — unequal cohort sizes, Gemma2's sparse Q2-2024 coverage ($n = 7$), and the restriction to four architecture families — are flagged explicitly in the limitations. The analysis is descriptive and observational; no causal claims are made without appropriate hedging. All citations were verified against the original papers before inclusion.

AI Tool Reflection

AI tools played a role at multiple stages of this project, supporting writing, research, and problem-solving along the way. They helped suggest structure, and accelerate the exploration of ideas that would have otherwise taken significantly longer to develop. All AI-generated content was reviewed critically and revised where needed. On several occasions, outputs were plausible but inaccurate or poorly suited to the specific context, requiring manual correction and judgment to align with the actual goals of the project. The most significant learning impact was recognizing that AI tools shift the work rather than eliminate it. The effort moves from generating content from scratch to evaluating, refining, and taking ownership of what is produced. That distinction matters. Going forward, these tools seem most useful as a starting point and thinking aid, but decisions about accuracy, relevance, and intent still require human judgment and cannot be delegated to the model.

Impact Statement

This work studies the psychometric stability of publicly available LLM benchmark data and introduces no new models, datasets, or deployment systems. The primary positive impact is methodological: clearer criteria for anchor-item selection could improve the validity of cross-generation LLM comparisons, benefiting researchers and practitioners who rely on leaderboard rankings for model selection decisions. A potential risk is misuse of the typology framework — specifically, using the "invariant" label to justify anchor selection without accounting for estimation uncertainty or the subtask-level aggregation limitation. We mitigate this by explicitly stating that `temporal_sequences` is the *only* universally stable BBH anchor and that item-level validation is needed before any anchor is used in production calibration. All data used is publicly available under HuggingFace's terms of service; no human participants, sensitive personal data, or proprietary information are involved. Attribution follows standard academic practice; all external code libraries are cited and licensed (Apache 2.0 for `torch_measure`).