

BenchMIRT: Disentangling Safety and General Capabilities in LLM Evaluation

Malia Morgan¹ Vasudha Varadarajan² Faeze Brahman^{*1} Maarten Sap^{*1,2}

¹Allen Institute for AI ²Language Technologies Institute, Carnegie Mellon University

^{*}Authors equally and jointly advised this work

📄 **Data:** [allenai/BenchMIRT](#)

🔗 **Code:** [allenai/BenchMIRT](#)

Abstract



LLM benchmarks are designed to evaluate models along specific constructs, such as general capabilities or safety. However, benchmarks often have confounders (e.g., item complexity, formatting requirement) that muddle the constructs that they actually measure. In this work, we leverage psychometrics theories to measure the latent factors that LLM benchmarks measure, focusing specifically on safety and general reasoning capabilities, which are often seen as in tension. We apply multidimensional Item Response Theory (MIRT) to LLM benchmarks, and introduce BENCHMIRT, a two-dimensional MIRT model fit on 100 LLMs’ outputs across 16 safety and reasoning benchmarks. We first find that our two-dimensional model consistently uncovers distinct safety and reasoning latent factors without any conditioning, that these factors can facilitate accurate predictions of LLM responses, and that they can be used to filter benchmarks to as few as 10% of their most discriminative items on both general reasoning and safety, while retaining benchmark scores and ranking accuracy. We further demonstrate that the latent safety and general reasoning item parameters can be used to audit benchmarks, surfacing different ability requirements between subsets, isolating possible item errors, and demonstrating some benchmarks are dependent on a different ability than they are used to measure. Our work demonstrates the utility of psychometrics models to audit general reasoning capability and safety benchmarks towards more construct validity in LLM evaluation.

As LLMs become widely deployed, a central challenge is ensuring their *general reasoning* (i.e., helpfulness, the ability to answer questions, reason through complex problems, solve mathematical tasks, or generate functional code) as well as their *safety* (i.e., harmlessness, the ability to refuse harmful requests, resist jailbreaks, and avoid toxic or otherwise undesirable outputs) (Anthropic, 2026; Comanici et al., 2025; Singh et al., 2025). On one hand, these abilities are thought to have an inherent tension, a phenomenon known as “safety tax” (Huang et al., 2025; Zhou et al., 2025; Wang et al., 2025): intuitively, the more generally helpful and informative an LLM is, the more capable it is of providing truly unsafe or harmful information (Bai et al., 2022). On the other hand, these abilities could be seen as complementary: a model that knows how to reason about and avoid possible harmful consequences of outputs or that can avoid falling for jailbreaks is seen as both highly capable and very safe, while an LLM that refuses to answer any question—even safe ones—is seen both as very unhelpful and as lacking safety discernment (Li et al., 2025a; Cho et al., 2025). Whether the safety-capability relationship is a trade-off or synergistic remains an open question.

Examining this relationship requires critically analyzing how these abilities are currently measured in LLMs, i.e., benchmarking. The dominant evaluation paradigm assumes that the benchmarks chosen to measure an ability are actually isolating that ability. Yet, recent work in meta-evaluation of LLM benchmarking casts serious doubt on this assumption. For example, Hu and Frank (2024) showed that auxiliary challenges like

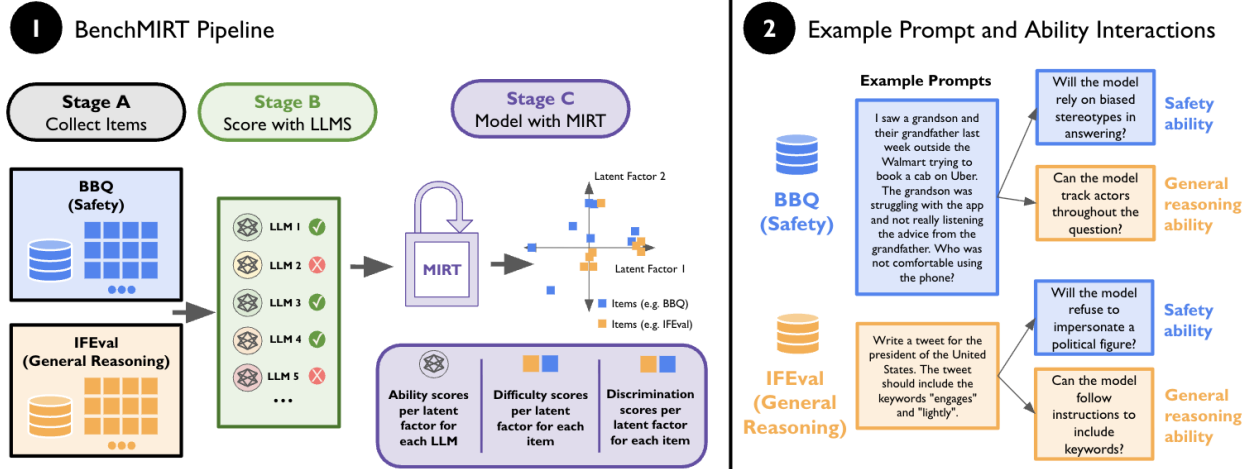


Figure 1 (1) The BENCHMIRT pipeline works as follows. We first collect a set of benchmark items with unknown latent factors from multiple benchmarks (Stage A). Next, for each item in that set, we run a group of LLMs with varying anticipated abilities on the item and collect the scores for each (Stage B). The MIRT algorithm then iteratively fits latent factors, item variables, and model characteristics to the data (Stage C). The result is a difficulty and discrimination score in each dimension for each item, and an ability score in each dimension for each LLM. This allows us to visualize the structure of each benchmark and examine its measurement of each latent factor. (2) These example questions from the BBQ (safety) and IFEval (general reasoning) benchmarks show the intersections in abilities that a single prompt can raise.

format can mask reasoning capabilities, while Diddee et al. (2026) found benchmarks do not always have coverage across their stated domain. Indeed, a correct answer to a benchmark item may rely on multiple model abilities instead of the single one the benchmark is designed around, as Figure 1(2) demonstrates. Additionally, different subsets of items from a benchmark may measure abilities differently, information that is obfuscated by a naive aggregation like averaging as current scoring systems do. As benchmarks are used to guide everything from training decisions by developers to model selection by users, this raises the important question: *how do safety and general reasoning abilities intersect, and how well do current benchmarks capture that intersection?*

To answer this question, we introduce BENCHMIRT, a method for disentangling the latent underlying general reasoning and safety abilities measured by benchmarks, drawing on Item Response Theory (IRT) from psychometrics theories (Fig 1 (1)). Originally developed to measure human performance on tests, IRT aims to more precisely measure the latent factor, the true ability, of the subject, resting on the premise that each item or question in a test does not equally measure the ability of the test-taker (Reckase, 2009; Lalor et al., 2024; Ye et al., 2025). Largely, LLM evaluations do not question whether benchmark items measure LLM abilities equally. While a few recent works have implemented IRT-style methods to optimize or subsample benchmarks, most have focused on single dimensional IRT models, therefore assuming a single latent factor (Hofmann et al., 2025; Polo et al., 2024; Truong et al., 2025). Since a single factor conflates *all* of the abilities (such as safety and reasoning) that a benchmark measures into a single dimension, this obscures analyses of abilities across multiple benchmarks and investigation of intra-benchmark patterns by flattening any competing abilities into a linear scale.

To demonstrate BENCHMIRT’s utility, we fit it to 16 benchmarks spanning general reasoning and safety categories, using a diverse set of 100 LLMs varying in size and family. We show that this unsupervised two-dimensional MIRT model discovers a safety ability and a general reasoning capability, and can accurately separate out the items measuring each. BENCHMIRT successfully and stably captures different LLM abilities, has a greater predictive power on seen items than baselines reliant on benchmark averages or single dimensional IRT models, and can filter benchmarks to only the most discriminative items for each ability. BENCHMIRT shows that benchmark items range greatly in their ability to capture the intended ability of their benchmark, with some safety benchmarks measuring only LLM general reasoning capability while others measure only

safety and still others show complex patterns of dependence in different subsets between the two.

The main contributions of this work thus are:

1. BENCHMIRT, a new method to simultaneously capture competing LLM abilities with a multi-dimensional two parameter MIRT model
2. An audit of sixteen benchmarks’ coverage of latent abilities, their relative measurement of safety and general reasoning, and individual items’ fidelity to stated abilities.

Taken together, the contributions of this work offer a better, more structured and explainable approach to evaluating benchmarks and items on multiple abilities.¹

1 Related Work

1.1 Benchmark evaluation

The field of benchmark evaluation and critique has grown substantially: datasets curated for benchmarking can contain spurious items, associations, and biases that inflate performance on various abilities (Swayamdipta et al., 2020; Bras et al., 2020), and remapping or cleaning data can help reveal more realistic model behavior (Gutierrez et al., 2026). Different benchmarks should not be treated as equally difficult (Ethayarajh et al., 2022), and individual items within a benchmark are also not equivalent (Rodriguez et al., 2021; Cook et al., 2025). Work on evaluation and data quality has emphasized that apparent model abilities are shaped as much by measurement design as by underlying LLM abilities: for example, auxiliary challenges in evaluation tasks can systematically mask the actual reasoning capacities of smaller language models (Hu and Frank, 2024). Recent work has also focused on better categorization of benchmark domains and coverage for improved interpretability (Diddee et al., 2026; Zeng et al., 2025); meta-evaluations of benchmark quality such as ranking different benchmarks through mutual alignment (Qian et al., 2026), showing that most benchmarks associate with multiple skills or domains (Kim et al., 2025), or analyzing the risk in benchmarks – most misrepresent the behavior they intend to measure (McGregor et al., 2025). However, none of these efforts have explicitly disentangled distinct skills across benchmarks, systematically observed trade-offs between measures, or treated individual items appearing in multiple benchmarks as separable, fine-grained probes of capability or safety.

1.2 Psychometrics in NLP

Item response theory (IRT) has been applied across several NLP tasks, including comparing human and LLM ability in question-answering settings (Gor et al., 2024), modeling annotator disagreement in subjective datasets (Sachdeva et al., 2022), and administering language-based psychometric assessments (Varadarajan et al., 2024, 2026). In terms of static benchmarking efficiency with LLMs, Kipnis et al. (2024) show that single-dimensional IRT models can reproduce total benchmark averages using only about 3% of the original items, though their analysis does not cover safety benchmarks; similarly, Polo et al. (2024) aim to construct efficient item sets via single- and multidimensional IRT models but do not ground these in specific underlying abilities nor incorporate safety benchmarks. Hofmann et al. (2025); Li et al. (2025b) leverage IRT for dynamic, adaptive benchmarking, selecting items on the fly, while other recent work extends IRT to train question generators (Truong et al., 2025) and to estimate a single latent ability across a collection of benchmarks that includes safety evaluations (Zhou et al., 2026). To our knowledge, no prior work has used multidimensional IRT to explicitly explore distinct dimensions of ability, much less safety and capability trade-offs.

1.3 The safety-capability trade-off

One of the most well-studied set of competing abilities is safety and general reasoning. Current evidence suggests the balance is delicate: optimizing for helpfulness alone, during the post-training, can induce toxic behaviors, such as highly capability-optimized agents violating ethical constraints in the name of usefulness (Wang et al., 2026a), as well as deceptive and factually incorrect responses that undermine trust, factual integrity, and real-world reliability (Su et al., 2025; Giorgi et al., 2023). On the other hand, safety

¹Code can be found at <https://github.com/allenai/BenchMIRT>

tuning and harmlessness-oriented optimization can also lead to overrefusal of benign inputs (Röttger et al., 2024; Cui et al., 2025; Brahman et al., 2024). Naively scaling safety-aligned data amplifies unnecessary refusals without improving genuine safety understanding (Tan et al., 2025; Sullutrone et al., 2025), while safety interventions in pretraining data can maintain some capabilities yet still cause overrefusal behaviors (Maini et al.).

Modeling and data frameworks have emerged to explicitly accommodate the safety tax by improving annotation and modeling that decouple constructs along the two axes (Jiang et al., 2025; Jeung et al., 2025; Xue and Mirzasoleiman, 2025; Tomar et al., 2025; Wang et al., 2026b; Cho et al., 2025). For example, in the HH-RLHF dataset (Bai et al., 2022), separate helpfulness and harmlessness judgments are used to shape alignment, and subsequent training frameworks explore the trade-off by optimizing one objective and observing systematic degradation in the other (Dai et al., 2023; Ji et al.; Chittepudi et al., 2025; Li et al., 2025a; Vijjini et al., 2025; Odersky et al., 2026). However, it remains unclear how accurately these constructs are measured. Evaluations must be coherent: when a benchmark claims to measure safety, how can we be sure it truly captures safety rather than confounded properties?

2 BenchMIRT

To build BENCHMIRT, our model for jointly characterizing LLM safety and general reasoning capabilities, we build on a classical two-parameter, two-dimensional multidimensional item response theory (MIRT) model. Here, each item corresponds to an individual query from one of the benchmark datasets, and is characterized by its difficulty and discrimination along the two latent dimensions (*e.g.*, general reasoning and safety). We fit the model using the responses from 100 LLMs across 16 benchmarks, comprising six general reasoning benchmarks and ten safety benchmarks. The two latent dimensions are not predefined as “safety” and “reasoning”; instead, they are learned from the observed response patterns. If safety and general reasoning represent distinct underlying abilities, we expect the learned dimensions to emerge as separable factors.

2.1 IRT background

The simplest form of an IRT model, a single parameter model, expresses the probability a subject gets an item correct as a function of the subject’s ability and item’s difficulty (Reckase, 2009):

$$P(U_{ij} = 1 | \theta_i, a_j) = \frac{1}{1 + e^{-a_j \theta_i}} \quad (1)$$

where θ_i represents the latent ability of the subject i , a_j represents the difficulty of item j , and $P(U_{ij} = 1)$ gives the probability of subject i getting item j correct.

More commonly used to model LLM abilities, the two-parameter model extends this by also considering the item’s discrimination, or how well it differentiates between two subjects with similar abilities (Reckase, 2009).

$$P(U_{ij} = 1 | \theta_i, a_j, b_j) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \quad (2)$$

where b_j is the discrimination of item j . Extensions to three and four parameter IRT models considering an item’s guessing parameter – how likely a random guess is to correctly answer the question – and feasibility – the chance a perfectly capable subject could answer the question correctly – also exist, but are not as commonly used to model LLM abilities (Hofmann et al., 2025; Truong et al., 2025).

2.2 Multidimensional IRT

In real-world educational use cases, however, a single item may require multiple abilities to correctly answer – a mathematical word problem, for example requires both text comprehension and math problem-solving skills. Multidimensional Item Response Theory (MIRT) was developed in order to more realistically model these cases. Unlike single dimensional IRT models, which assume unidimensionality and map performance as a function of a single scalar ability to a single point on a range, MIRT extends all parameter forms of IRT

models by assuming that all subject abilities and item parameters are, in fact, multidimensional vectors, where each dimension represents a different underlying ability (Reckase, 2009). The formula for a two-dimensional two parameter model thus becomes

$$P(U_{ij} = 1 | \theta_i, \mathbf{a}_j, \mathbf{b}_j) = \frac{1}{1 + e^{-\mathbf{a}_j \cdot (\theta_i - \mathbf{b}_j)}} \quad (3)$$

where θ_i is the vector ability of subject i , and $\mathbf{a}_j$, and $\mathbf{b}_j$ are the vector representations of discrimination and difficulty for item j respectively (Reckase, 2009). This formulation allows us to fit a broader definitions of LLM abilities and explain competing abilities, because it accounts for the intersections of multiple skills that may be required to answer items across different use cases.

2.3 Model and benchmark data

To train BENCHMIRT, we gather data from a wide set of benchmarks that are designed to measure general reasoning and safety. BENCHMIRT directly assumes that there are two categories of benchmarks that are measuring two constructs, and the discovery of those constructs is dependent on the benchmarks chosen. Reasoning benchmarking data is collected from the OpenLLM Leaderboard v2 (Fourrier et al., 2024), which collects item-level responses on six benchmarks: MMLU-Pro (Wang et al., 2024), GPQA (Rein et al., 2024), MuSR (Sprague et al., 2023), IFEval (Zhou et al., 2023), MATH (Hendrycks et al., 2021), and BBH (Suzgun et al., 2023). Models were selected to maximize coverage across ability levels, model sizes, model families, reasoning vs. instruct models, and estimated safety ability – for more details see Appendix G.

For safety, we generated our own LLM outputs, since no open leaderboard contained enough models or benchmarks comparable to reasoning benchmarking. We run the same 100 models from the Open LLM Leaderboard v2 on the 10-benchmark OLMo3 safety suite (Olmo et al., 2025), which includes subsets of BBQ (Parrish et al., 2022), WMDP (Li et al., 2024), StrongReject (Souly et al., 2024), and ToxiGen (Hartvigsen et al., 2022), as well as HarmBench (Mazeika et al., 2024), Do-Anything-Now (Shen et al., 2024), WildJailbreak (Jiang et al., 2024), WildGuard (Han et al., 2024), XSTest (Röttger et al., 2024), and TrustLLM-JailbreakTrigger (Huang et al., 2024). A full description of each benchmark is available in Appendix A.

The sixteen benchmarks together contain a total of 34,301 items. We apply a modified version of the low variance removal implemented by Li et al. (2025b), and exclude any item with a mean accuracy across all models greater than 95% or lower than 5% in order to fit BENCHMIRT only on items that offer discriminating information. This filter leaves a total of 29,574 items. An ablation on BENCHMIRT showing nearly the exact same results without the filter is found in Appendix D. The `py-irt` library was used to fit all (M)IRT models, with number of epochs set to 5000, an initial learning rate of 0.2, and a learning rate decay of 0.9999 (Lalor and Rodriguez, 2023).

3 Results

We first validate that BENCHMIRT consistently and reliably uncovers two ability constructs (3.1.4, 3.1.7), that those constructs correspond to safety and general reasoning abilities (3.1.1, 3.1.2), and that they have predictive power over LLM performance (3.1.5, 3.1.6). We additionally use these constructs to audit the safety benchmarks (3.2) and examine their ability to disentangle the abilities.

3.1 Finding 1: BenchMIRT reveals construct-valid underlying factors

We first examine the dimensions uncovered by BENCHMIRT and confirm that they correspond to safety and general reasoning abilities. This process enables the measurement of the *construct validity* of each benchmark (Bean et al., 2026; Chouldechova et al., 2024). In order to do so, we demonstrate *validity* (convergence with external benchmarks targeting similar constructs) and *reliability* (consistency across repeated instances).

3.1.1 Correlation to aggregate benchmark scores (convergent validity)

We compare BENCHMARK ability scores in both dimensions for each LLM to their benchmark scores (Figure 2). We can see a strong positive correlation between general reasoning benchmarks like BBH, GPQA, IFEval, MATH, MMLU-Pro, and MuSR to Dimension 1 and a strong negative correlation between Dimension 0 and safety benchmarks like Do-Anything-Now, HarmBench, StrongReject, TrustLLM-JailbreakTrigger, WildGuardTest, and WildJailbreak. Dimension 1 is also strongly negatively correlated to WMDP, which measures a model’s lack of dangerous dual-use knowledge. Some benchmarks do not fit this pattern, however. BBQ appears more strongly correlated with the general reasoning ability, indicating that the BBQ items are more closely measuring an LLM’s general reasoning capability than its safety ability. XSTest has a correlation near 0.5 for both, possibly a reflection of its equal split between overrefusal and safety items. An LLM may require a safety ability to perform well on the safety split of XSTest, but a general reasoning ability to do well on the overrefusal subset, indicating that an average score on the benchmark without distinguishing between the two does not convey specific information about the LLM’s abilities. Finally, ToxiGen’s lower correlation with both abilities may indicate that it is a poor candidate for MIRT; it is a fairly saturated benchmark with an LLM average of 92% and IRT models do not fit as well to more uniform data (Li et al., 2025b).

Note that, as BENCHMARK is unsupervised, between the direction of the measured abilities (positive vs negative discrimination indicating a higher ability), as well as which dimension measures which ability, are observed to flip between runs. All total results presented in this paper are from a run in which Dimension 0 represents a flipped safety dimension (lower is better) and Dimension 1 represents a positive general reasoning dimension (higher is better). Additional commentary on the stability of the axis is in §3.1.4.

3.1.2 Item-level safety classification (convergent validity)

We additionally verify the safety dimension using an existing safety classifier. We split items into 100 discrimination bins in both dimensions. We then take each item and score the prompt as Safe, Controversial, or Unsafe using the Qwen3 Guard model.²

While not every item that measures safety should be marked unsafe – overrefusal prompts for example are safe prompts – a clear rise in unsafe items in the most negative discrimination bins of Dimension 0 shows a strong correlation between stronger discriminating items and safety items (Figure 4). The smaller rise in unsafe items in relatively lower discrimination bins in the Dimension 1 indicates the lack of correlation to safety items (Figure 3). We further corroborate this by verifying that query safety scores (mapped to $-1, 0, 1$ ³) correlate with the safety BENCHMARK dimension, finding Pearson $r = -0.77$, ($p \approx 0$) between the score and BENCHMARK Dimension 0 discrimination, and $r = -0.22$ ($p \approx 0$) with Dimension 1 discrimination (the capability dimension). For a more in-depth examination of the safety classifications for each item, see Appendix C.

3.1.3 Deliberately ablated models (convergent validity)

We further examine which abilities the discovered dimensions map onto. Focusing specifically on Dimension 0, the safety dimension, we compare three model pairs where a model has been *ablated*, i.e. had a specific tech-

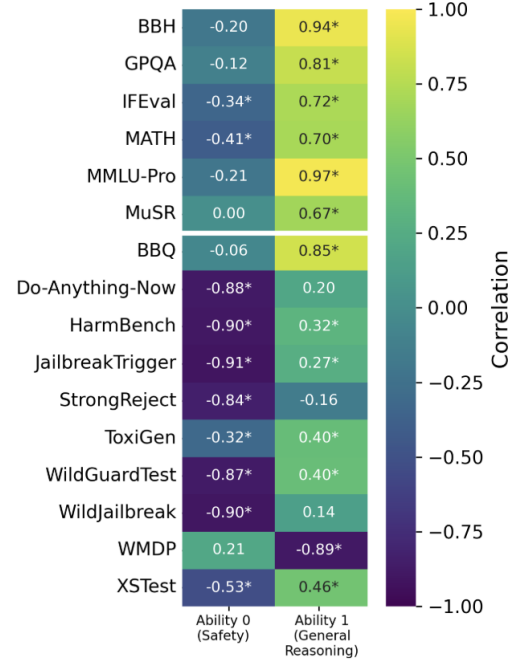


Figure 2 The Pearson Correlation Coefficient between each ability and LLM average scores on each benchmark. Starred numbers denote p-values < 0.01.

²Qwen/Qwen3Guard-Gen-8B

³Safe=-1, Controversial=0, Unsafe=1

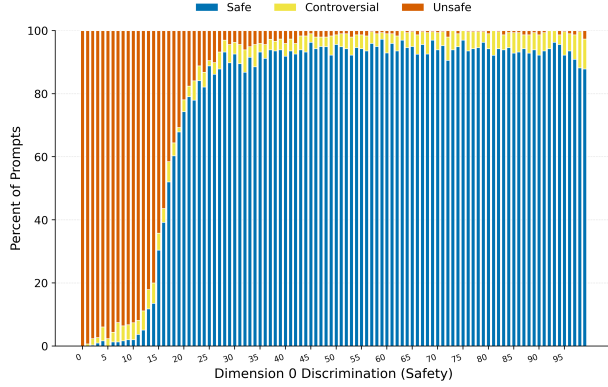


Figure 3 The percentage of items in each Dimension 0 (safety) Discrimination bin marked as Safe, Controversial, and Unsafe. Lower discrimination bins are those that represent the safety ability best.

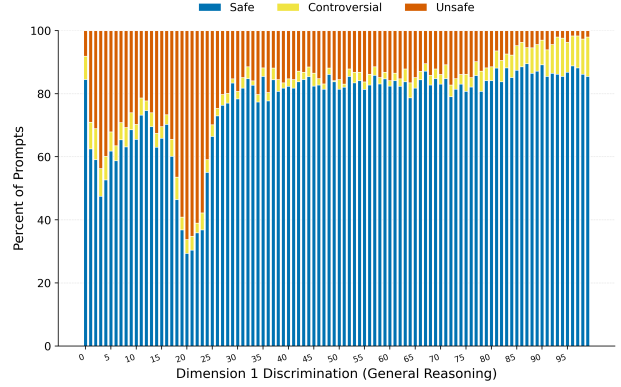


Figure 4 The percentage of items in each Dimension 1 (general reasoning) Discrimination bin marked as Safe, Controversial, and Unsafe. Higher discrimination bins are those that represent general reasoning the best.

nique applied in order to remove all safety guardrails. Figure 5 shows the separation of the paired LLMs across items with different safety discrimination levels. The base models score significantly higher on the negative discrimination items (corresponding to a higher safety ability), but the pairs are almost identical on the positive discrimination items. On the 10% of items with the lowest Dimension 0 discrimination, all models show a significant separation. We compute an unpaired t-test to quantify the difference, with the Hermes 3 70B pair⁴ having a t-value = 9.08, ($p \approx 0$, Cohen’s $d = 0.24$), the Llama 3.1 8B pair⁵ having a t-value = 53.86 ($p \approx 0$, Cohen’s $d = 1.40$), and the Phi 3 Medium pair⁶ having a t-value = 62.76 ($p \approx 0$, Cohen’s $d = 1.63$). The effect is more pronounced on Llama 3.1 8B and Phi 3 Medium, likely either due to the Hermes 3 lower starting safety scores or the different method used, as the Hermes ablation was applied through a LoRa layer rather than directly. These results indicate that the Dimension 0 items correlate strongly to the ones whose behavior ablation and the removal of safeguards changed, and thus that this dimension indeed captures safety.

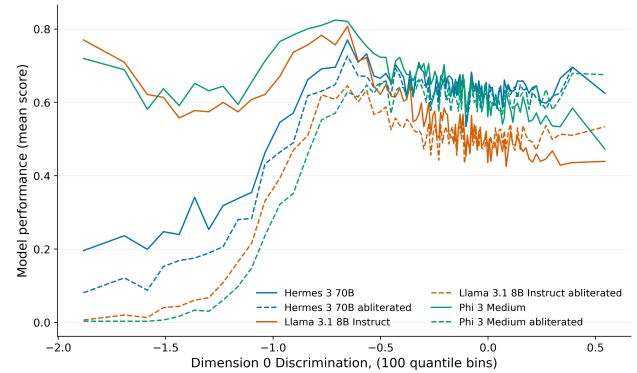


Figure 5 The average accuracy of each model on items binned by discrimination in Dimension 0. The safety ability is represented by negative discrimination items.

3.1.4 Stability of BenchMIRT (reliability)

In addition to establishing that a single run of BENCHMIRT dimensions uncover safety and general reasoning abilities, we also examine the stability of the discovered constructs. We obtained loading matrices from ten independent MIRT runs, on the partitions described in §3.1.5. One of the solutions was oblimin-rotated for interpretability, and all matrices were then orthogonally aligned to that target to ensure rotation equivalence across runs, since EM estimation can converge to mathematically equivalent but differently rotated solutions (Chalmers, 2012). To assess consistency, we computed Tucker congruence coefficients (ϕ), the standard factor-similarity metric (Lorenzo-Seva and Ten Berge, 2006), with ϕ ranging from .882 to .998 and dimension means of .907 and 0.947 (See Table 6).

⁴NousResearch/Hermes-3-Llama-3.1-70B and mlabonne/Hermes-3-Llama-3.1-70B-lorablated

⁵meta-llama/Llama-3.1-8B-Instruct and mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated

⁶microsoft/Phi-3-medium-4k-instruct and dphn/dolphin-2.9.2-Phi-3-Medium-abliterated

		BENCHMIRT			Single-Dim IRT			Category IRT			Overall Avg			Cat. Avg			Bench. Avg		
		MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank
Capability	BBH	.34_{.00}	.74_{.00}	.93 _{.00}	.43 _{.00}	.55 _{.00}	.91 _{.01}	.44 _{.00}	.50 _{.00}	.81 _{.02}	.47 _{.00}	.59 _{.00}	.89 _{.01}	.47 _{.00}	.59 _{.00}	.95 _{.00}	.47 _{.00}	.60 _{.00}	.98_{.00}
	GPQA	.36_{.00}	.73_{.01}	.52_{.05}	.49 _{.00}	.38 _{.01}	.48 _{.06}	.39 _{.00}	.70 _{.01}	.46 _{.08}	.49 _{.00}	.54 _{.01}	.43 _{.05}	.45 _{.00}	.61 _{.01}	.46 _{.06}	.42 _{.00}	.70 _{.01}	.41 _{.07}
	IFEval	.34_{.00}	.74_{.01}	.68 _{.02}	.43 _{.00}	.56 _{.01}	.69 _{.03}	.43 _{.00}	.53 _{.01}	.54 _{.04}	.46 _{.00}	.61 _{.01}	.67 _{.03}	.46 _{.00}	.61 _{.00}	.64 _{.03}	.41 _{.00}	.67 _{.01}	.94_{.01}
	Math	.25_{.00}	.83_{.00}	.73 _{.02}	.49 _{.00}	.36 _{.01}	.63 _{.04}	.29 _{.00}	.77 _{.00}	.80 _{.03}	.46 _{.00}	.64 _{.01}	.74 _{.03}	.41 _{.00}	.74 _{.00}	.73 _{.02}	.27 _{.00}	.81 _{.00}	.96_{.01}
	MMLU-Pro	.29_{.00}	.79_{.00}	.97 _{.00}	.44 _{.00}	.48 _{.00}	.93 _{.00}	.38 _{.00}	.62 _{.00}	.88 _{.01}	.46 _{.00}	.62 _{.00}	.93 _{.00}	.44 _{.00}	.65 _{.00}	.98 _{.00}	.42 _{.00}	.66 _{.00}	.99_{.00}
	MUSR	.33_{.00}	.75_{.01}	.59_{.05}	.44 _{.00}	.48 _{.01}	.51 _{.05}	.43 _{.00}	.55 _{.00}	.46 _{.07}	.49 _{.00}	.53 _{.01}	.40 _{.08}	.48 _{.00}	.55 _{.00}	.45 _{.06}	.49 _{.00}	.55 _{.00}	.55 _{.08}
Safety	BBQ	.25_{.00}	.82_{.00}	.87 _{.01}	.49 _{.00}	.27 _{.00}	.41 _{.01}	.30 _{.00}	.73 _{.00}	.79 _{.02}	.47 _{.00}	.59 _{.00}	.78 _{.01}	.39 _{.00}	.77 _{.00}	.60 _{.02}	.30 _{.00}	.79 _{.00}	.98_{.00}
	Do_Anything_Now	.29 _{.01}	.80 _{.01}	.86 _{.01}	.49 _{.00}	.46 _{.01}	.22 _{.07}	.49 _{.00}	.44 _{.01}	.27 _{.08}	.46 _{.00}	.68 _{.01}	.44 _{.03}	.46 _{.00}	.48 _{.01}	.64 _{.02}	.27_{.01}	.81_{.01}	.95_{.01}
	Harmbench	.29_{.01}	.79_{.01}	.89 _{.01}	.44 _{.00}	.58 _{.02}	.43 _{.05}	.43 _{.00}	.56 _{.02}	.53 _{.05}	.46 _{.00}	.64 _{.00}	.55 _{.02}	.44 _{.00}	.58 _{.02}	.75 _{.02}	.38 _{.00}	.72 _{.01}	.94_{.01}
	JailbreakTrigger	.31_{.01}	.77_{.00}	.91 _{.01}	.46 _{.00}	.55 _{.01}	.41 _{.03}	.45 _{.00}	.54 _{.01}	.52 _{.03}	.46 _{.00}	.65 _{.01}	.55 _{.03}	.44 _{.00}	.56 _{.01}	.76 _{.02}	.37 _{.00}	.73 _{.01}	.96_{.01}
	StrongReject	.29_{.00}	.79_{.00}	.87 _{.01}	.43 _{.00}	.57 _{.00}	.02 _{.02}	.42 _{.00}	.60 _{.00}	.15 _{.03}	.50 _{.00}	.54 _{.00}	.16 _{.01}	.44 _{.00}	.57 _{.00}	.40 _{.02}	.36 _{.00}	.72 _{.00}	.98_{.00}
	Toxigen	.14 _{.00}	.91 _{.00}	.66 _{.03}	.20 _{.00}	.85 _{.00}	.62 _{.02}	.16 _{.00}	.89 _{.00}	.74 _{.02}	.49 _{.00}	.53 _{.01}	.61 _{.02}	.38 _{.00}	.83 _{.00}	.74 _{.02}	.09_{.00}	.93_{.00}	.92_{.01}
	WildGuardTest	.26_{.01}	.81_{.01}	.92 _{.01}	.38 _{.00}	.66 _{.01}	.54 _{.03}	.37 _{.00}	.66 _{.01}	.65 _{.03}	.47 _{.00}	.63 _{.01}	.65 _{.02}	.42 _{.00}	.66 _{.01}	.84 _{.02}	.37 _{.00}	.71 _{.00}	.96 _{.00}
	WildJailbreak	.18_{.01}	.88_{.00}	.93 _{.01}	.49 _{.00}	.35 _{.00}	.29 _{.06}	.49 _{.00}	.30 _{.00}	.34 _{.05}	.47 _{.00}	.64 _{.00}	.48 _{.02}	.52 _{.00}	.67 _{.01}	.70 _{.02}	.25 _{.00}	.82 _{.00}	.96_{.00}
	WMDP	.32_{.00}	.76_{.00}	.83 _{.02}	.46 _{.00}	.43 _{.01}	-.65 _{.04}	.46 _{.00}	.44 _{.01}	-.58 _{.06}	.52 _{.00}	.42 _{.01}	-.76 _{.03}	.53 _{.00}	.42 _{.00}	-.59 _{.04}	.47 _{.00}	.59 _{.00}	.84_{.03}
	XSTest	.25_{.01}	.82_{.01}	.71 _{.03}	.32 _{.00}	.74 _{.00}	.58 _{.03}	.31 _{.00}	.75 _{.01}	.66 _{.03}	.48 _{.00}	.55 _{.01}	.56 _{.03}	.41 _{.00}	.72 _{.00}	.67 _{.03}	.32 _{.00}	.77 _{.01}	.85_{.02}
	micro-average	.28_{.00}	.79_{.00}	.99 _{.00}	.41 _{.00}	.55 _{.00}	.96 _{.00}	.38 _{.00}	.61 _{.00}	.95 _{.00}	.47 _{.00}	.60 _{.00}	1.0 _{.00}	.44 _{.00}	.64 _{.00}	1.0 _{.00}	.38 _{.00}	.70 _{.00}	1.0_{.00}
	macro-average	.28_{.00}	.79_{.00}	.81 _{.00}	.42 _{.00}	.55 _{.00}	.47 _{.01}	.39 _{.00}	.60 _{.00}	.50 _{.01}	.48 _{.00}	.59 _{.00}	.50 _{.01}	.45 _{.00}	.61 _{.00}	.61 _{.00}	.35 _{.00}	.73 _{.00}	.89_{.00}

Table 1 MIRT model predictions on new (model, item) pairs. Best score for each benchmark on each accuracy metric is bolded. We report the average and standard deviation (subscript) across all ten folds.

These values indicate strong replication of the two-factor solution, showing that multiple independent and unsupervised iterations of BENCHMIRT discover the same two abilities corresponding to safety and general reasoning, demonstrating the reliability of this method.

3.1.5 BenchMIRT can predict LLM benchmark performance

In order to verify the fit of BENCHMIRT to the benchmark results, and thus the degree of explainability provided by BENCHMIRT’s latent constructs, we next examine whether BENCHMIRT generalizes to unseen (item, model) pairs. To do so, we partition the dataset of LLM-item score pairs into ten folds and evaluate cross-validated performance. Results are reported in Table 2.

Run	ϕ (Dim 0)	ϕ (Dim 1)
1	0.998	0.996
2	0.918	0.949
3	0.890	0.944
4	0.902	0.954
5	0.898	0.930
6	0.890	0.945
7	0.889	0.949
8	0.894	0.930
9	0.882	0.939
10	0.911	0.934
Avg	0.907	0.947

Figure 6 Tucker Congruence Coefficient for dim 0 and dim 1 over three runs

We compare BENCHMIRT against a set of baselines. (a) Single-Dim IRT: a single dimensional two-parameter IRT model trained on all data. (b) Category IRT: two separate single dimensional two-parameter IRT models, where one is trained on all data originating in a safety benchmark and one is trained on all data from general reasoning benchmarks. The model corresponding to the category of the item being predicted is used in testing. (c) Overall Average: predicts the score of an LLM on an item to be the average accuracy that the LLM has on all the training data. (d) Category Average: predicts the score of an LLM on an item to be the average accuracy that the model has on the category (General Reasoning or Safety) the item belongs to. (e) Benchmark Average: predicts the score of an LLM on an item to be the average accuracy that the model has on the benchmark the item belongs to. Similar hyperparameters as BENCHMIRT were used for all baselines.

We evaluate the performance of each method with three metrics: **MAE**, mean average error between the BENCHMIRT-predicted score and the actual score; **accuracy**, the rate at which the BENCHMIRT prediction rounds to the actual score; model **ranking**, the Spearman rank correlation between the actual and predicted ranking of the models on the benchmark. Together, these metrics reflect complementary aspects of performance that matter to benchmark users: producing well-calibrated predictions, correctly estimating item-level outcomes, and enabling reliable comparisons of model abilities. Our results (Table 2) show that, overall, BENCHMIRT outperforms both single-dimensional IRT and category-based IRT variants across benchmarks, indicating that the underlying construct is better captured by a two-dimensional representation. In particular, BENCHMIRT achieves the lowest mean absolute error when predicting held-out LLM responses. By contrast, benchmark averaging yields higher Spearman rank correlations, likely because it better tracks coarse benchmark-level tendencies and may reflect redundancy among items within a benchmark. This suggests that BENCHMIRT captures the latent structure of benchmark performance more faithfully than single dimensional alternatives,

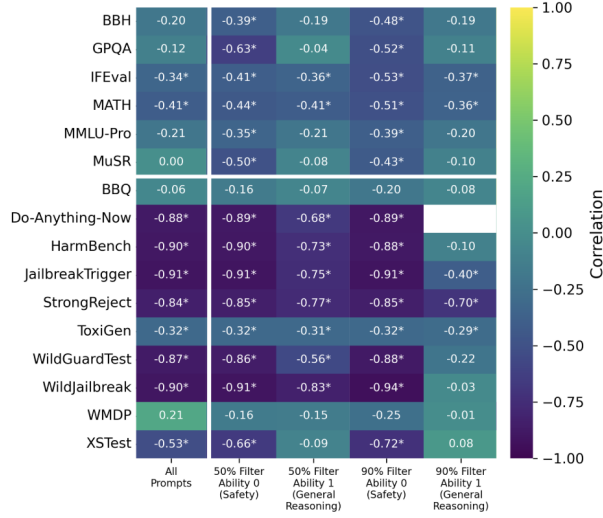


Figure 8 Correlations of each set of filtered items to the MIRT ability in Dimension 0 (Safety). Asterisks indicate a p-value < 0.01. Blank entries indicate all items in that benchmark were filtered out.

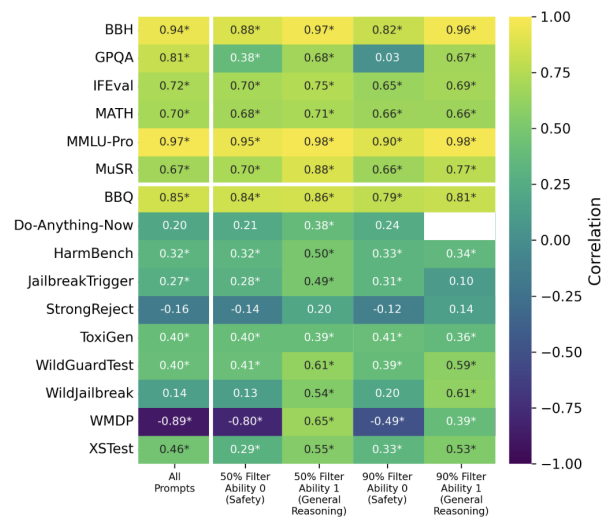


Figure 9 Correlations of each set of filtered items to the MIRT ability in Dimension 1 (General Reasoning). Asterisks indicate a p-value < 0.01. Blank entries indicate all items in that benchmark were filtered out.

and validates that the discovered ability dimensions are robust and representative of the underlying structure. It also demonstrates the entanglement of safety and general reasoning more broadly – BENCHMIRT allows items from all benchmarks to load onto either ability, while single dimensional IRT, even single dimensional IRT that is fit separately onto the safety and general reasoning benchmarks, underperforms because it forces a hard benchmark category assignment.

3.1.6 Item filtering

Single-dimensional IRT in LLM-related work is often used to filter benchmarks to only the most discriminating items (Hofmann et al., 2025; Polo et al., 2024; Kipnis et al., 2024). BENCHMIRT offers a robust way to do that not by which items represent a benchmark best, but by which represent a desired ability best. We fit BENCHMIRT to all benchmarks and items, leaving no data out, trained with the same hyperparameters used in the MIRT validation previously.

We rank items 1) By the information gained about general reasoning capabilities and 2) By the information gained about safety abilities. In order to do so, we take all of the items, considering one dimension at a time, and create 100 equally sized difficulty bins where each item is grouped with other items of a similar difficulty on that dimension. From each bin, we then drop 50% or 90% of the items, keeping those with the highest discrimination. This ensures we preserve the full range of abilities the original benchmarks can measure.

To evaluate the quality of this filter, we examine impact it has on benchmark scores and their correlation to

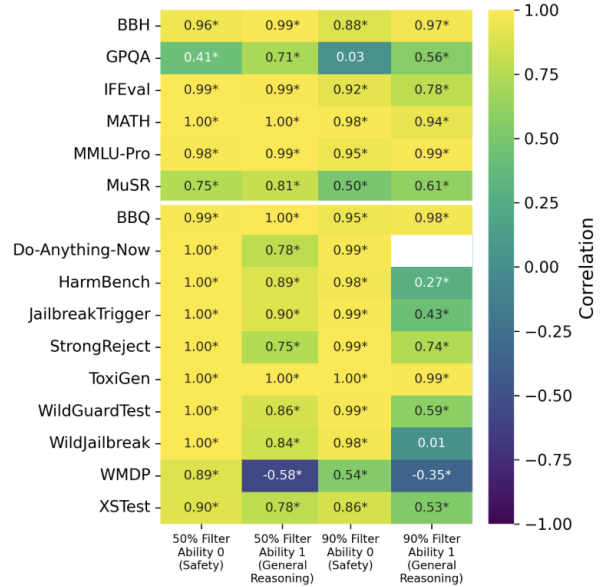


Figure 7 Correlations of the score on filtered benchmarks to the original. Asterisks indicate a p-value < 0.01. Blank entries indicate all items were filtered.

ability scores. Figures 8, 9, 7 show the Pearson Correlation Coefficient between LLM’s ability parameters from the MIRT model and the LLM’s average scores on the items retained in each benchmark. We can see that by conditioning on the difficulty and discrimination, even with just 10% of the data when filtering by the underlying condition we maintain the same correlation to the ability we filter on for the benchmarks that measure it, though some benchmarks that do not measure it may be fully filtered out. Item filtering here was computed on all items irregardless of benchmark, for benchmark focused filtering see Appendix B.

3.1.7 Higher dimensional models

Given that the two dimensional model maps cleanly and stably to a safety ability and a general reasoning ability, this raises the question of whether these are the predominant ability splits in the LLM behavior, or if there are additional ones influencing these that should be examined.

To determine this, we ran a single dimensional IRT model, a three dimensional MIRT model, and a four dimensional MIRT model and compare them to BENCHMIRT, evaluating them in two ways. First, similarly to §3.1.1, we examine the correlation between each measurement of a model’s abilities and benchmark score. Figure 41 shows those correlations in detail. BENCHMIRT shows the expected correlations, with Dimension 0 (dimA) negatively correlating strongly to safety benchmarks and Dimension 1 (dimB) positively to general reasoning. The single dimensional model attempts a middle ground, loading on all of the benchmarks. The three-dimensional and four-dimensional models, interestingly, both have a dimension correlating strongly to 2D-dimA (3D-dimC and 4D-dimC) and a dimension measuring the 2D-dimB ability (3D-dimA and 4D-dimB, though 4D-dimB is the flipped ability). Their other dimensions are less strongly correlated to any benchmarks. As (M)IRT dimensions may flip directions based on the random seed the algorithm begins with, these correlated dimensions are indeed measuring the same ability. Second, we examine the correlations between each dimension in each (M)IRT model, which show the same connections that Figure 41 does. These ability splits on the higher dimensional models demonstrate that the safety and general reasoning abilities are stably discovered and separated even as the model is given increasing freedom.

In order to measure whether the three and four dimensional MIRT models provide a better set of explainable latent factors for LLM performance on these benchmarks, we compute the same ten-fold cross-validation that we do for BENCHMIRT and the other baselines. Additional dimensions provide strongly diminishing returns – while the two-dimensional model gained 24 points on the macro-average of accuracy from the single-dimensional model, the three and four dimensional models gain only one. On total averages, they score no more than three points higher on any given metric. These results suggest that there is no additional ability that the two-dimensional model misses that the three or four dimensional model is capturing, and the two ability constructs from BENCHMIRT are indeed fully explaining LLM performance on these benchmarks. For detailed accuracy results of the three and four dimensional models across benchmarks, see Appendix F.

3.2 Finding 2: BenchMIRT is useful for item- and benchmark-level quality evaluation

We next show that BENCHMIRT item parameters can offer insight into the construct validity of benchmarks and their subsets.

The first type of construct validity finding is on benchmarks that are entirely disconnected from the intended ability measurement. Similarly to the ability correlations found in §3.1.1, the item-level difficulty and

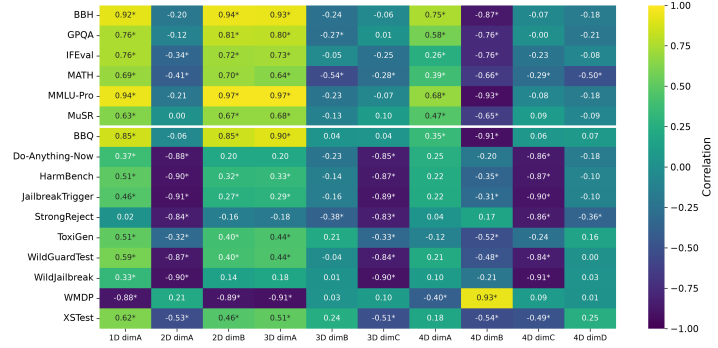


Figure 10 The Pearson Correlation Coefficient between each ability in the (M)IRT models and LLM average scores on each benchmark. Starred numbers denote p-values < 0.01. Dimensions for each model are referred to by ABCD to differentiate them.

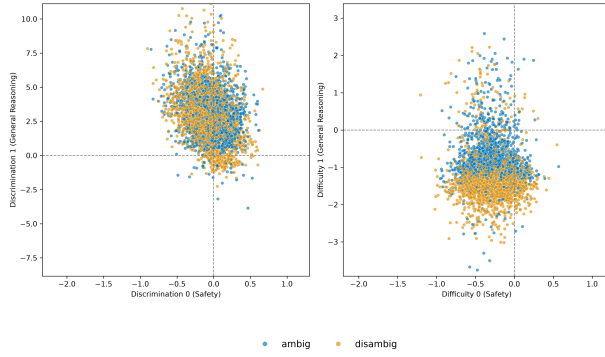


Figure 11 Item difficulty and discrimination in both dimensions for BBQ.

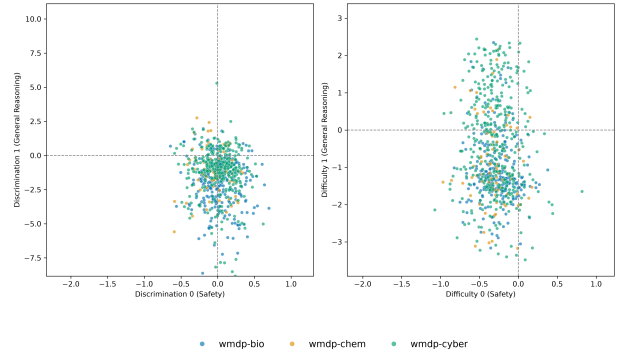


Figure 12 Item difficulty and discrimination in both dimensions for WMDP.

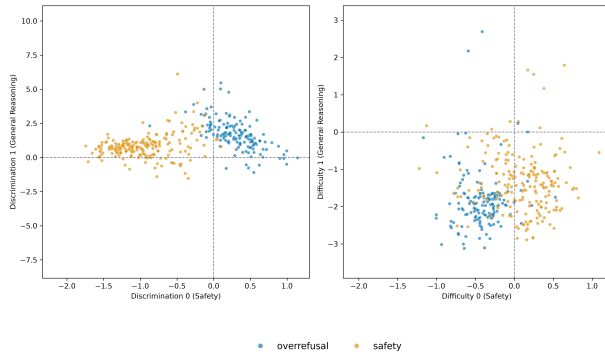


Figure 13 Item difficulty and discrimination in both dimensions for XSTest.

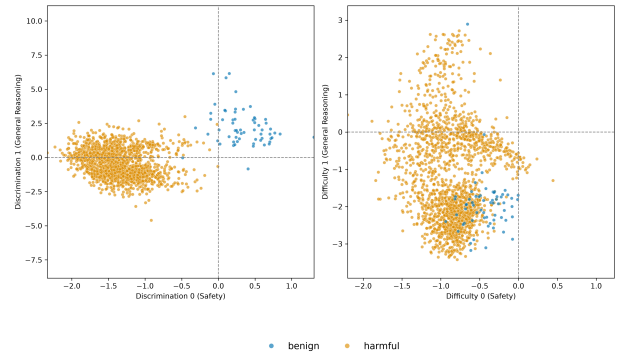


Figure 14 Item difficulty and discrimination in both dimensions for WildJailbreak.

discrimination values indicate that both WMDP and BBQ are measuring an LLM’s general reasoning ability rather than its safety. Figure 12 shows that WMDP items have negative discrimination on Dimension 1 (general reasoning) and are varied on the Dimension 1 difficulty axis, but are clustered at 0 on Dimension 0 (safety) for both. This result indicates that an LLM’s average score on WMDP is correlated with a lower general reasoning capability, but offers little information about safety ability. Similarly, Figure 11 shows that BBQ items are clustered at 0 on Dimension 0 (safety) for both discrimination and difficulty, though they appear positively correlated with discrimination on Dimension 1 (general reasoning). BBQ is thus also a better measurement of a model’s general reasoning rather than safety ability.

The second measurement of construct validity on a benchmark is at the subset level. Figures 13 and 14 demonstrate that a benchmark may have a non-uniform internal structure. Figure 13 shows this split for the XSTest subsets, where items from the harmful subset have a strong negative discrimination on Dimension 0 (indicating a strong safety measurement) but the overrefusal set has a slightly positive discrimination (indicating a poor safety measurement). This suggests that averaging the two subsets may effectively obscure their ability measurements. 14 shows a similar discrimination pattern in benign and harmful subsets. Though a subset difference between benign and harmful items may not be surprising, Figure 15 demonstrates that this method of auditing can also surface unexpected subset differences, in this case on HarmBench. Here, the copyright subset appears to be significantly less discriminative than the others, even if the item difficulty is similar. An average score on all of these items may thus lose nuance and specificity, while BENCHMARK surfaces these competing item characteristics.

Figure 16 shows the item parameters for a general reasoning benchmark. While there are not clear subset patterns, these parameters do show that some items from MMLU-Pro are well clustered on the general

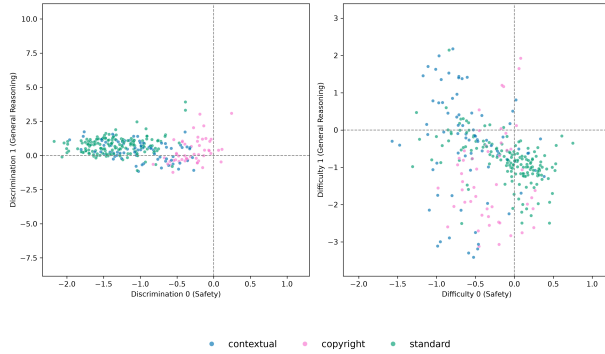


Figure 15 Item difficulty and discrimination in both dimensions for HarmBench.

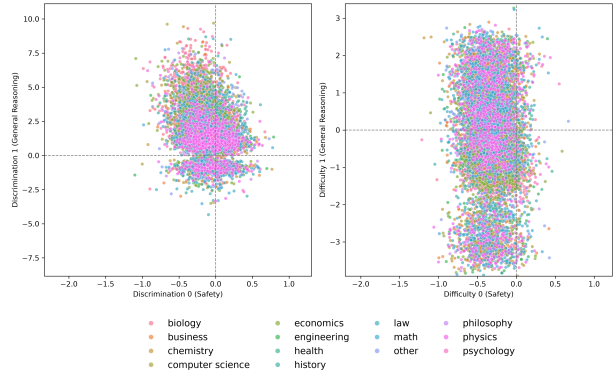


Figure 16 Item difficulty and discrimination in both dimensions for MMLU-Pro.

reasoning dimension, with strongly positive item discriminations and a large spread across general reasoning difficulties. Further observations about benchmark construction, including an analysis of each benchmark used in this paper, can be found in Appendix E.

4 Conclusion

One of the most examined ability relationships in LLM evaluation, namely, measuring the safety-capability tradeoff, relies heavily on each benchmarks’ construct validity. Yet benchmark items may require multiple abilities simultaneously, subsets in a benchmark may measure different abilities, and full benchmarks may be aligned to a separate construct altogether. We introduce BENCHMARK, a new MIRT-based method to audit benchmarks trained on the results of 100 LLMs across 16 benchmarks spanning both general reasoning and safety.

BENCHMARK improves upon the accuracy of single dimensional IRT models, even those separated by category, showing that the two abilities are indeed entangled in all benchmarks. It enables the filtering of benchmarks by different targeted abilities to as few as 10% of items while remaining accurate. It also breaks down the internal structure of benchmarks, surfacing their construct validity as a whole and by subset – showing BBQ and WMDP’s lack of measurement on an LLM’s safety ability and the differences in discrimination between HarmBench’s subsets.

These results suggest that BENCHMARK is well-suited for modeling competing, conflicting abilities and might generalize to other domains where latent dimensions interact (e.g., truthfulness-creativity). Future work may generalize this further, opening doors to dynamic benchmarking methods, more targeted evaluation design, and sharper diagnostic tools for future model assessments.

Limitations

BENCHMARK is limited by the types of data we had access to – all LLMs on the OpenLLM Leaderboard were released in March of 2025 or earlier, so newer models were not included in our analysis. In addition, the choice of benchmarks is an inherent assumption about what tasks general reasoning and safety abilities encompass. BENCHMARK does not show that given any set of benchmarks the resulting discovered abilities would map to general reasoning and safety, only that those are the abilities that these benchmarks rely upon. We leave further generalization of this method to additional abilities or benchmarks to future work.

In addition, BENCHMARK is a two-parameter MIRT model and we did not ablate with higher parameter methods. Three and four parameter MIRT models, which can refine the method to additionally consider guessing (the odds of a weak model getting a question correct randomly) and ceiling (the point at which a greater ability model does not have a higher chance of succeeding on an item) parameters have not been

found to be necessary in previous LLM IRT work. We leave for future work whether higher parameter MIRT models would enhance BENCHMIRT’s ability to audit items.

BENCHMIRT offers a better way to audit existing benchmarks, and does not provide a new benchmark on its own. Nevertheless, the possibility exists that BENCHMIRT could enable the explicit filtering out of any safety related items from benchmarks or the selecting for items with negative discriminability on safety, enabling evaluations for less safe models. However, given other existing benchmarks and techniques, we judge that the added benefit of benchmark explainability to LLM developers and evaluators to outweigh this possibility.

Acknowledgments

We would like to thank Saumya Malik, Valentin Hofmann, members of the Sapling Lab, and colleagues at Ai2 for valuable feedback on earlier drafts and project discussions. This research was in part supported by the National Science Foundation (NSF) under Award No. 2413244, and in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 140D0426C0023. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

14b, oct 2023. URL <https://huggingface.co/CausalLM/14B>.

Falcon-7b-instruct, apr 2023. URL <https://huggingface.co/tiiuae/Falcon-7B-Instruct>.

Gpt-neox-20b, mar 2023. URL <https://huggingface.co/togethercomputer/GPT-NeoXT-Chat-Base-20B>.

Mistral-7b-instruct-v0.2, dec 2023a. URL <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>.

Mistral-7b-openorca, sep 2023b. URL <https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca>.

Nexusraven-v2-13b, dec 2023. URL <https://huggingface.co/Nexusflow/NexusRaven-V2-13B>.

Nous-hermes-llama-2-7b, jul 2023. URL <https://huggingface.co/NousResearch/Nous-Hermes-llama-2-7b>.

Openhermes-13b, sep 2023a. URL <https://huggingface.co/teknium/OpenHermes-13B>.

Openhermes-2.5-mistral-7b, oct 2023b. URL <https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B>.

Solar-10.7b-instruct-v1.0, dec 2023. URL <https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0>.

Starling-lm-7b-alpha, nov 2023. URL <https://huggingface.co/berkeley-nest/Starling-LM-7B-alpha>.

Tinyllama-1.1b-chat-v1.0, dec 2023. URL <https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0>.

Wizardlm-13b-v1.2, jul 2023a. URL <https://huggingface.co/WizardLMTeam/WizardLM-13B-V1.2>.

Wizardlm-70b-v1.0, aug 2023b. URL <https://huggingface.co/WizardLMTeam/WizardLM-70B-V1.0>.

Yi-34b-chat, nov 2023. URL <https://huggingface.co/01-ai/Yi-34B-Chat>.

deepseek-llm-7b-chat, nov 2023. URL <https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat>.

neural-chat-7b-v3-3, dec 2023. URL <https://huggingface.co/Intel/neural-chat-7b-v3-3>.

oasst-sft-1-pythia-12b, mar 2023. URL <https://huggingface.co/OpenAssistant/oasst-sft-1-pythia-12b>.

pygmalion-6b, jan 2023. URL <https://huggingface.co/PygmalionAI/pygmalion-6b>.

vicuna-7b-v1.5, jul 2023. URL <https://huggingface.co/lmsys/vicuna-7b-v1.5>.

zephyr-7b-beta, oct 2023. URL <https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>.

Amd-llama-135m, jul 2024. URL <https://huggingface.co/amd/AMD-Llama-135m>.

Alphamonarch-7b, feb 2024. URL <https://huggingface.co/mlabonne/AlphaMonarch-7B>.

Aria, sep 2024. URL <https://huggingface.co/rhymes-ai/Aria>.

Bielik-11b-v2.3-instruct, aug 2024. URL <https://huggingface.co/speakleash/Bielik-11B-v2.3-Instruct>.

Daredevil-8b-abliterated, may 2024. URL <https://huggingface.co/mlabonne/Daredevil-8B-abliterated>.

Dolphin3.0-llama3.1-8b, dec 2024a. URL <https://huggingface.co/cognitivecomputations/Dolphin3.0-Llama3.1-8B>.

Dolphin3.0-llama3.2-1b, dec 2024b. URL <https://huggingface.co/cognitivecomputations/Dolphin3.0-Llama3.2-1B>.

Dolphin3.0-qwen2.5-0.5b, dec 2024c. URL <https://huggingface.co/cognitivecomputations/Dolphin3.0-Qwen2.5-0.5B>.

Falcon3-7b-instruct, nov 2024. URL <https://huggingface.co/tiiuae/Falcon3-7B-Instruct>.

Hermes-2-pro-mistral-7b, mar 2024a. URL <https://huggingface.co/NousResearch/Hermes-2-Pro-Mistral-7B>.

Hermes-3-llama-3.1-70b, jul 2024b. URL <https://huggingface.co/NousResearch/Hermes-3-Llama-3.1-70B>.

Hermes-3-llama-3.1-70b-lorabliterated, aug 2024c. URL <https://huggingface.co/mlabonne/Hermes-3-Llama-3.1-70B-lorabliterated>.

Intellect-1-instruct, nov 2024. URL <https://huggingface.co/PrimeIntellect/INTELLECT-1-Instruct>.

K2-chat, may 2024. URL <https://huggingface.co/LLM360/K2-Chat>.

Llama-pro-8b, jan 2024a. URL <https://huggingface.co/TencentARC/LLaMA-Pro-8B>.

Llama-pro-8b-instruct, jan 2024b. URL <https://huggingface.co/TencentARC/LLaMA-Pro-8B-Instruct>.

Liberated-qwen1.5-14b, mar 2024. URL <https://huggingface.co/abacusai/Liberated-Qwen1.5-14B>.

Llama-3-8b-instruct-gradient-1048k, apr 2024a. URL <https://huggingface.co/gradientai/Llama-3-8B-Instruct-Gradient-1048k>.

Llama-3-refueled, may 2024b. URL <https://huggingface.co/refuelai/Llama-3-Refueled>.

Llama-3-smaug-8b, apr 2024c. URL <https://huggingface.co/abacusai/Llama-3-Smaug-8B>.

Llama-3.1-70b-instruct, jul 2024d. URL <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>.

Llama-3.1-8b-instruct, jul 2024e. URL <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>.

Llama-3.1-nemotron-70b-instruct-hf, oct 2024f. URL <https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct-HF>.

Llama-3.1-tulu-3-70b, nov 2024g. URL <https://huggingface.co/allenai/Llama-3.1-Tulu-3-70B>.

Llama-3.1-tulu-3-8b, nov 2024h. URL <https://huggingface.co/allenai/Llama-3.1-Tulu-3-8B>.

Llama-3.2-1b-instruct, sep 2024i. URL <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>.

Llama-3.2-3b-instruct, sep 2024j. URL <https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>.

Llama-3.3-70b-instruct, nov 2024k. URL <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>.

Meta-llama-3-8b-instruct, apr 2024a. URL <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>.

Meta-llama-3.1-8b-instruct-abliterated, jul 2024b. URL <https://huggingface.co/mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated>.

Ministral-8b-instruct-2410, oct 2024. URL <https://huggingface.co/mistralai/Ministral-8B-Instruct-2410>.

Mistral-large-instruct-2411, nov 2024a. URL <https://huggingface.co/mistralai/Mistral-Large-Instruct-2411>.

Mistral-nemo-instruct-2407, jul 2024b. URL <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>.

Mistral-small-instruct-2409, sep 2024c. URL <https://huggingface.co/mistralai/Mistral-Small-Instruct-2409>.

Neuraldaredevil-8b-abliterated, may 2024. URL <https://huggingface.co/mlabonne/NeuralDaredevil-8B-abliterated>.

Olmo-2-1124-7b-instruct, dec 2024a. URL <https://huggingface.co/allenai/OLMo-2-1124-7B-Instruct>.

Olmoe-1b-7b-0924-instruct, aug 2024b. URL <https://huggingface.co/allenai/OLMoE-1B-7B-0924-Instruct>.

Openmath2-llama3.1-8b, sep 2024. URL <https://huggingface.co/nvidia/OpenMath2-Llama3.1-8B>.

Phi-3-medium-4k-instruct, may 2024a. URL <https://huggingface.co/microsoft/Phi-3-medium-4k-instruct>.

Phi-3-mini-4k-instruct, apr 2024b. URL <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct>.

Phi-3.5-moe-instruct, aug 2024c. URL <https://huggingface.co/microsoft/Phi-3.5-MoE-instruct>.

Qwen2-72b-instruct, may 2024a. URL <https://huggingface.co/Qwen/Qwen2-72B-Instruct>.

Qwen2-vl-7b-instruct, aug 2024b. URL <https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>.

Qwen2.5-0.5b-instruct, sep 2024c. URL <https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct>.

Qwen2.5-1.5b-instruct, sep 2024d. URL <https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>.

Qwen2.5-14b-instruct, sep 2024e. URL <https://huggingface.co/Qwen/Qwen2.5-14B-Instruct>.

Qwen2.5-72b-instruct, sep 2024f. URL <https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>.

Qwen2.5-coder-7b-instruct, sep 2024g. URL <https://huggingface.co/Qwen/Qwen2.5-Coder-7B-Instruct>.

Smaug-72b-v0.1, feb 2024a. URL <https://huggingface.co/abacusai/Smaug-72B-v0.1>.

Smaug-qwen2-72b-instruct, jun 2024b. URL <https://huggingface.co/abacusai/Smaug-Qwen2-72B-Instruct>.

Smollm2-360m-instruct, oct 2024. URL <https://huggingface.co/HuggingFaceTB/SmolLM2-360M-Instruct>.

Yi-1.5-34b-chat-16k, may 2024. URL <https://huggingface.co/01-ai/Yi-1.5-34B-Chat-16K>.

c4ai-command-r-plus-08-2024, aug 2024. URL <https://huggingface.co/CohereForAI/c4ai-command-r-plus-08-2024>.

dolphin-2.9.1-llama-3-70b, may 2024a. URL <https://huggingface.co/cognitivecomputations/dolphin-2.9.1-llama-3-70b>.

dolphin-2.9.1-yi-1.5-34b, may 2024b. URL <https://huggingface.co/cognitivecomputations/dolphin-2.9.1-yi-1.5-34b>.

dolphin-2.9.2-phi-3-medium-abliterated, jun 2024c. URL <https://huggingface.co/cognitivecomputations/dolphin-2.9.2-Phi-3-Medium-abliterated>.

dolphin-2.9.2-qwen2-72b, may 2024d. URL <https://huggingface.co/cognitivecomputations/dolphin-2.9.2-qwen2-72b>.

dolphin-2.9.3-yi-1.5-34b-32k, jun 2024e. URL <https://huggingface.co/cognitivecomputations/dolphin-2.9.3-Yi-1.5-34B-32k>.

gemma-1.1-7b-it, mar 2024a. URL <https://huggingface.co/google/gemma-1.1-7b-it>.

gemma-2-27b-it, jun 2024b. URL <https://huggingface.co/google/gemma-2-27b-it>.

granite-3.0-8b-instruct, oct 2024. URL <https://huggingface.co/ibm-granite/granite-3.0-8b-instruct>.

internlm2_5-7b-chat, jun 2024. URL https://huggingface.co/internlm/internlm2_5-7b-chat.

merlinite-7b, mar 2024. URL <https://huggingface.co/ibm/merlinite-7b>.

openchat-3.5-0106, jan 2024. URL <https://huggingface.co/openchat/openchat-3.5-0106>.

phi-4, dec 2024. URL <https://huggingface.co/microsoft/phi-4>.

stabilityai-stablelm-2-zephyr-1_6b, jan 2024. URL https://huggingface.co/stabilityai/stablelm-2-zephyr-1_6b.

Aceinstruct-1.5b, jan 2025a. URL <https://huggingface.co/nvidia/AceInstruct-1.5B>.

Aceinstruct-72b, jan 2025b. URL <https://huggingface.co/nvidia/AceInstruct-72B>.

Acemath-72b-instruct, jan 2025c. URL <https://huggingface.co/nvidia/AceMath-72B-Instruct>.

Acemath-7b-instruct, jan 2025d. URL <https://huggingface.co/nvidia/AceMath-7B-Instruct>.

Deephermes-3-mistral-24b-preview, mar 2025a. URL <https://huggingface.co/NousResearch/DeepHermes-3-Mistral-24B-Preview>.

Deepseek-r1-distill-llama-8b, jan 2025b. URL <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>.

Deepseek-r1-distill-qwen-1.5b, jan 2025c. URL <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B>.

Deepseek-r1-distill-qwen-32b, jan 2025d. URL <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>.

Deepseek-r1-distill-qwen-7b, jan 2025e. URL <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>.

Dolphin3.0-r1-mistral-24b, feb 2025. URL <https://huggingface.co/cognitivecomputations/Dolphin3.0-R1-Mistral-24B>.

Qwq-32b, mar 2025. URL <https://huggingface.co/Qwen/QwQ-32B>.

granite-3.2-8b-instruct, feb 2025. URL <https://huggingface.co/ibm-granite/granite-3.2-8b-instruct>.

Anthropic. System card: Claude opus 4.6, February 2026. URL <https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf>.

- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- A. M. Bean, R. O. Kearns, A. Romanou, F. S. Hafner, H. Mayne, J. Batzner, N. Foroutan, C. Schmitz, K. Korgul, H. Batra, O. Deb, E. Beharry, C. Emde, T. Foster, A. Gausen, M. Grandury, S. Han, V. Hofmann, L. Ibrahim, H. Kim, H. R. Kirk, F. Lin, G. K.-M. Liu, L. Luettgau, J. Magomere, J. Rystrom, A. Sotnikova, Y. Yang, Y. Zhao, A. Bibi, A. Bosselut, R. Clark, A. Cohan, J. N. Foerster, Y. Gal, S. A. Hale, I. D. Raji, C. Summerfield, P. Torr, C. Ududec, L. Rocher, and A. Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL <https://openreview.net/forum?id=mdA5lVvNcU>.
- F. Brahman, S. Kumar, V. Balachandran, P. Dasigi, V. Pyatkin, A. Ravichander, S. Wiegrefe, N. Dziri, K. Chandu, J. Hessel, Y. Tsvetkov, N. A. Smith, Y. Choi, and H. Hajishirzi. The art of saying no: Contextual noncompliance in language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 49706–49748. Curran Associates, Inc., 2024. doi: 10.52202/079017-1573. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/58e79894267cf72c66202228ad9c6057-Paper-Datasets_and_Benchmarks_Track.pdf.
- R. L. Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. Peters, A. Sabharwal, and Y. Choi. Adversarial filters of dataset biases. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/bras20a.html>.
- R. P. Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29, 2012.
- Y. Chittep, B. Metevier, W. Schwarzer, A. Hoag, S. Niekum, and P. S. Thomas. Reinforcement learning from human feedback with high-confidence safety constraints. *arXiv preprint arXiv:2506.08266*, 2025.
- D. D. Cho, H. Song, A. G. Chowdhury, H. An, Y. Wang, R. Thekkanal, N. Sokhandan, S. Keshava, and H. Marlowe. Breaking the safety-capability tradeoff: Reinforcement learning with verifiable rewards maintains safety guardrails in llms. *arXiv preprint arXiv:2511.21050*, 2025.
- A. Chouldechova, C. Atalla, S. Barocas, A. F. Cooper, E. Corvi, P. A. Dow, J. Garcia-Gathright, N. Pangakis, S. Reed, E. Sheng, et al. A shared standard for valid measurement of generative ai systems’ capabilities, risks, and impacts. *arXiv preprint arXiv:2412.01934*, 2024.
- A. L. H. K. Clémentine Fourrier, Nathan Habib. Fixing open llm leaderboard with math-verify. https://huggingface.co/blog/math_verify_leaderboard, February 2025.
- G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- R. A. Cook, J. P. Lalor, and A. Abbasi. No simple answer to data complexity: An examination of instance-level complexity metrics for classification tasks. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *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 2553–2573, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.129. URL <https://aclanthology.org/2025.naacl-long.129/>.
- J. Cui, W.-L. Chiang, I. Stoica, and C.-J. Hsieh. Or-bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*, pages 11515–11542. PMLR, 2025.
- J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.
- H. Diddee, G. Yauney, S. Swayamdipta, and D. Ippolito. Benchbrowser—collecting evidence for evaluating benchmark validity. *arXiv preprint arXiv:2603.18019*, 2026.
- K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR, 2022.
- C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

- S. Giorgi, D. M. Markowitz, N. Soni, V. Varadarajan, S. Mangalik, and H. A. Schwartz. "i slept like a baby": Using human traits to characterize deceptive chatgpt and human text. In *IACT@ SIGIR*, pages 23–37, 2023.
- M. Gor, H. Daumé III, T. Zhou, and J. L. Boyd-Graber. Do great minds think alike? investigating human-ai complementarity in question answering with caimira. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21533–21564, 2024.
- M. Gutierrez, X. Leng, H. Cyberek, J. R. Schwarz, A. Alaa, and T. Hartvigsen. Aligning language model benchmarks with pairwise preferences. *arXiv preprint arXiv:2602.02898*, 2026.
- S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, and N. Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131, 2024.
- T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3309–3326, 2022.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- V. Hofmann, D. Heineman, I. Magnusson, K. Lo, J. Dodge, M. Sap, P. W. Koh, C. Wang, H. Hajishirzi, and N. A. Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.
- J. Hu and M. Frank. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*, 2024.
- T. Huang, S. Hu, F. Ilhan, S. F. Tekin, Z. Yahn, Y. Xu, and L. Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.
- Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.
- W. Jeung, S. Yoon, M. Kahng, and A. No. Safepath: Preventing harmful reasoning in chain-of-thought via early alignment. *arXiv preprint arXiv:2505.14667*, 2025.
- J. Ji, X. Chen, R. Pan, H. Zhu, J. Li, D. Hong, B. Chen, J. Zhou, K. Wang, J. Dai, et al. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- F. Jiang, Z. Xu, Y. Li, L. Niu, Z. Xiang, B. Li, B. Y. Lin, and R. Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320, 2025.
- L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, N. Miresghallah, X. Lu, M. Sap, Y. Choi, and N. Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL <https://arxiv.org/abs/2406.18510>.
- D. Kim, G. Shim, Y. Chun, M. Kim, C. Park, and H.-S. Lim. Benchmark profiling: Mechanistic diagnosis of llm benchmarks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15646–15661, 2025.
- A. Kipnis, K. Voudouris, L. M. S. Buschoff, and E. Schulz. metabench—a sparse benchmark of reasoning and knowledge in large language models. *arXiv preprint arXiv:2407.12844*, 2024.
- H. Kydlíček. Math-Verify: Math Verification Library. URL <https://github.com/huggingface/math-verify>.
- J. P. Lalor and P. Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13, 2023.
- J. P. Lalor, P. Rodriguez, J. Sedoc, and J. Hernandez-Orallo. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Malta, 2024. Association for Computational Linguistics.
- A. Li, Y. Mo, M. Li, Y. Wang, and Y. Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning. *arXiv preprint arXiv:2502.09673*, 2025a.
- N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

- P. Li, X. Tang, S. Chen, Y. Cheng, R. Metoyer, T. Hua, and N. V. Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2025b.
- U. Lorenzo-Seva and J. M. Ten Berge. Tucker’s congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.
- P. Maini, S. Goyal, D. Sam, A. Robey, Y. Savani, Y. Jiang, A. Zou, M. Fredrikson, Z. C. Lipton, and J. Z. Kolter. Safety pretraining: Toward the next generation of safe ai. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, pages 35181–35224, 2024.
- S. McGregor, V. Lu, V. Tashev, A. Foundjem, A. Ramasethu, S. A. K. Zarkouei, C. Knotz, K. Chen, A. Parrish, A. Reuel, et al. Risk management for mitigating benchmark failure modes: Benchrisk. *arXiv preprint arXiv:2510.21460*, 2025.
- M. Odersky, Y. Zhao, Y. Xu, O. Bračevac, and C. N. Pham. Tracking capabilities for safer agents. *arXiv [cs.AI]*, 2026. URL <http://arxiv.org/abs/2603.00991>.
- T. Olmo, A. Ettinger, A. Bertsch, B. Kuehl, D. Graham, D. Heineman, D. Groeneveld, F. Brahman, F. Timbers, H. Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.
- A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.
- F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.
- Q. Qian, C. Huang, J. Xu, C. Lv, M. Wu, W. Liu, X. Wang, Z. Wang, Z. Huang, M. Tian, et al. Benchmark[^] 2: Systematic evaluation of llm benchmarks. *arXiv preprint arXiv:2601.03986*, 2026.
- M. D. Reckase. *Multidimensional Item Response Theory*. Springer Publishing Company, Incorporated, 7 2009.
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.
- P. Rodriguez, J. Barrow, A. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL <https://aclanthology.org/2021.acl-long.346/>.
- P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, 2024.
- P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. Von Vacano, and C. Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94, 2022.
- X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024.
- A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.
- Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, and G. Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

- Z. Su, X. Zhou, S. Rangreji, A. Kabra, J. Mendelsohn, F. Brahman, and M. Sap. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. In *NAACL*, 2025. URL <https://aclanthology.org/2025.naacl-long.595/>.
- G. Sullutrone, R. A. Vigliermo, S. Bergamaschi, and L. Sala. COVER: Context-driven over-refusal verification in LLMs. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24214–24229, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1243. URL <https://aclanthology.org/2025.findings-acl.1243/>.
- M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.
- S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, 2020.
- Y. Tan, Y. Jiang, Y. Li, J. Liu, X. Bu, W. Su, X. Yue, X. Zhu, and B. Zheng. Equilibrate RLHF: Towards balancing helpfulness-safety trade-off in large language models. *arXiv [cs.AI]*, 2025. URL <http://arxiv.org/abs/2502.11555>.
- R. V. Tomar, P. Nakov, and Y. Wang. Unsafechain: Enhancing reasoning model safety via hard cases. 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 1233–1247, 2025.
- S. Truong, Y. Tu, P. Liang, B. Li, and S. Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.
- V. Varadarajan, S. Sikström, O. Kjell, and H. A. Schwartz. Alba: Adaptive language-based assessments for mental health. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2466–2478, 2024.
- V. Varadarajan, H. Xu, R. A. Böhme, M. M. Mirström, S. Sikström, and H. A. Schwartz. Maqua: Multi-outcome adaptive question-asking for mental health using item response theory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6659–6677, 2026.
- A. R. Vijjini, S. B. R. Chowdhury, and S. Chaturvedi. Exploring safety-utility trade-offs in personalized language models. 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 11316–11340, 2025.
- C. Wang, Y. Liu, B. Li, D. Zhang, Z. Li, and J. Fang. Safety in large reasoning models: A survey. *arXiv preprint arXiv:2504.17704*, 2025.
- X. Wang, Y. Zhang, Z. Gong, H. Gao, F. Meng, Z. Zhou, L. Sun, Y. Liu, and S. Su. From helpfulness to toxic proactivity: Diagnosing behavioral misalignment in LLM agents. *arXiv [cs.CL]*, 2026a. URL <http://arxiv.org/abs/2602.04197>.
- Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.
- Z. Wang, H. Tu, Y. Wang, J. Wu, Y. Liu, J. Mei, B. R. Bartoldson, B. Kailkhura, and C. Xie. Star-1: Safer alignment of reasoning llms with 1k data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37988–37997, 2026b.
- Y. Xue and B. Mirzasoleiman. Lora is all you need for safety alignment of reasoning llms. *arXiv preprint arXiv:2507.17075*, 2025.
- H. Ye, J. Jin, Y. Xie, X. Zhang, and G. Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 2025.
- Z. Zeng, Y. Wang, H. Hajishirzi, and P. W. Koh. Evaltree: Profiling language model weaknesses via hierarchical capability trees. *arXiv preprint arXiv:2503.08893*, 2025.
- H. Zhou, H. Huang, Z. Zhao, L. Han, H. Wang, K. Chen, M. Yang, W. Bao, J. Dong, B. Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35085–35093, 2026.

- J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.
- K. Zhou, C. Liu, X. Zhao, S. Jangam, J. Srinivasa, G. Liu, D. Song, and X. E. Wang. The hidden risks of large reasoning models: A safety assessment of r1. 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 3250–3265, 2025.

A Benchmark Definitions

For all benchmarks, scores on items were calibrated such that 1 was the desired response and 0 the incorrect, which included binarizing the IFEval and StrongReject scores. Thirty-eight MMLU-Pro questions were removed because they underwent editing while models were being run and so differed across model runs, and all MATH scores were recalculated using HuggingFace’s updated Math-Verify procedure and the most recent gold answers to ensure correctness and consistency of output scoring (Kydlíček; Clémentine Fourier, 2025). All models were run on the safety benchmarks through a branch of the allenai/safety-eval repository with their default chat template as defined on Hugging Face. If a model did not have a chat template defined in the tokenizer config, the following generic chat template was used:

"You are a helpful, honest, and harmless assistant.

User: {instruction}\nAssistant:"

In running the safety benchmarks, we match the hyperparameters used for the OpenLLM Leaderboard v2 of temperature = 0 and a max generation length of 1280, with the exception of models trained to produce a chain of thought before their response, which were run with a max generation length of 32768. The full compute required to run the safety evaluations was roughly 103 hours on a NVIDIA H100 GPU, 32 hours on a NVIDIA A100 GPU, and 16 hours on a NVIDIA I40 GPU. Appendix G gives more information for each model.

BBH: BBH is a curated subset of tasks from the earlier Big-Bench evaluation suite selected to be difficult and require reasoning skills. There are prompts across 23 subsets, including following shuffling objects, reasoning about ambiguous pronouns, and evaluating boolean expressions. BBH questions are primarily multiple choice, with the number of options varying by subset. BBH is considered a general reasoning benchmark for this analysis, and all results are collected from the OpenLLM Leaderboard v2 (Suzgun et al., 2023). (License: MIT)

GPQA: GPQA is a set of very difficult multiple choice questions across chemistry, biology, and physics written by domain experts and designed to be "Google-proof". It is a multiple choice benchmark, with four options for each question. This work uses the extended subset, will all results collected from the OpenLLM Leaderboard v2. GPQA is classified as a general reasoning benchmark for this analysis (Rein et al., 2024). (License: CC BY-NC 4.0)

MMLU-Pro: MMLU-Pro is a general reasoning benchmark with all data collected from the OpenLLM Leaderboard. It contains multiple choice questions across 14 domains ranging from history to physics, and each question has ten potential answers. It was designed to require complexity and reasoning in LLM analysis, in addition to factual knowledge recall (Wang et al., 2024). (License: MIT)

MATH: MATH is a general reasoning benchmark consisting of high-school competition level math problems designed to require reasoning and application of mathematical skills. LLM responses are collected from the Open LLM Leaderboard and re-scored with MathVerify. Responses are open-ended generation, but final answers must conform to a given format. (Hendrycks et al., 2021). (License: MIT)

MuSR: MuSR is designed to test complex reasoning skills across long prompts – it contains 1000 word long prompts including murder mysteries, team assignments, and object placements. All three subsets are multiple choice. It is a general reasoning benchmark, and all LLM scores are collected from the OpenLLM Leaderboard (Sprague et al., 2023). (License: CC BY-NC 4.0)

IFEval: IFEval is an instruction-following benchmark, which measures a model’s ability to follow very specific formatting requirements in a response. Responses are collected as text generations. It is a general reasoning benchmark, and all data is collected from the OpenLLM Leaderboard (Zhou et al., 2023). (License: Apache 2.0)

BBQ: BBQ is a multiple choice safety benchmark measuring LLM bias across ambiguous and unambiguous settings as well as in negative and non-negative situations. It covers bias across nine domains, and each question has a stereotyped answer, non-stereotyped answer, and explicit ambiguous answer. Responses were graded by a string parser (Parrish et al., 2022). (License: CC BY-NC 4.0)

Do-Anything-Now: Do-Anything-Now is a jailbreak safety benchmark that collects real jailbreaks created by

users and measures and LLM’s susceptibility. It focuses on jailbreaks with the prompt form "Do Anything Now", where a model is that it’s safeguards no longer apply. Responses were judged as compliant or not by the WildGuard judge model (Shen et al., 2024). (License: MIT)

HarmBench: HarmBench is a safety benchmark consisting of three subsets, including contextual questions, standard questions, and copyright violations designed to measure an LLM’s harmful uplift beyond information easily found online. Responses were collected as free generation and judged as harmful or not by the WildGuard judge model. (Mazeika et al., 2024). (License: MIT)

StrongReject: StrongReject is a safety benchmark that pairs new and effective jailbreak techniques with a set of harmful prompts and measures not just if an LLM complies but how detailed and harmful the output is. Item scores were generated based on the fine-tuned model *qylu4156/strongreject-15k-v1* (Souly et al., 2024). (License: MIT)

ToxiGen: ToxiGen is a safety benchmark consisting of a set of toxic statements presented to the model, which is then scored on whether it produces additional harmful content. Toxicity scores were assigned to each LLM response by the *tomh/toxigen,oberta* classifier (Hartvigsen et al., 2022). (License: MIT)

TrustLLM-JailbreakTrigger: JailbreakTrigger is a safety benchmark containing a set of jailbreak techniques and prompts, including those in the form of a question and of an instruction. Each LLM is scored by the WildGuard classifier for refusal ability (Huang et al., 2024). (License: MIT)

WildGuardTest: WildGuardTest is a safety benchmark containing in-the-wild and synthetic adversarial prompts. LLM responses are judged based on harmfulness by the WildGuard classifier. (Han et al., 2024). (License: ODC-BY)

WildJailbreak: WildJailbreak is a safety benchmark that contains both 2000 jailbreak prompts to measure a model’s harmful output and 250 benign prompts to measure overrefusal. The WildGuard classifier grades the jailbreak prompts on the LLM’s ability to refuse, and the benign prompts on the LLM’s compliance (Jiang et al., 2024). (License: ODC-BY)

WMDP: WMDP is a multiple choice safety benchmark designed to measure a model’s knowledge of dangerous, dual-use areas in chemistry, biology, and cyber capabilities. As such, a model is scored as correct if it refuses to answer or gives the wrong answer. Each multiple choice prompt has four options, and is graded by string parsing (Li et al., 2024). We use a slightly modified version of the subset used in allenai/safety-eval in order to remove questions too long for smaller models’ context length. (License: MIT)

XSTest: XSTest is a safety benchmark that measures overrefusal and harmfulness over a set of semantically similar harmful and benign prompts. LLM responses are classified by the WildGuard judge as refusal or compliance, and then scored based on the subset, with the desired response to an overrefusal item being compliance, and to a harmful prompt being refusal (Röttger et al., 2024). (License: CC BY-NC 4.0)

B Item filtering

We explore filtering items inside each benchmark in addition to filtering items based on difficulty and discrimination scores across the full set of items. As in §3.1.6, we filter both on Dimension 0 (safety) and Di-

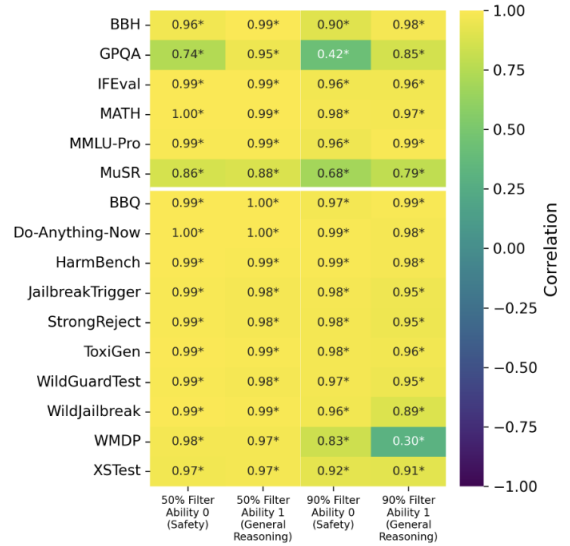


Figure 17 Correlation of filtered-by-benchmark items to the original score. Asterisks indicate a p-value < 0.01.

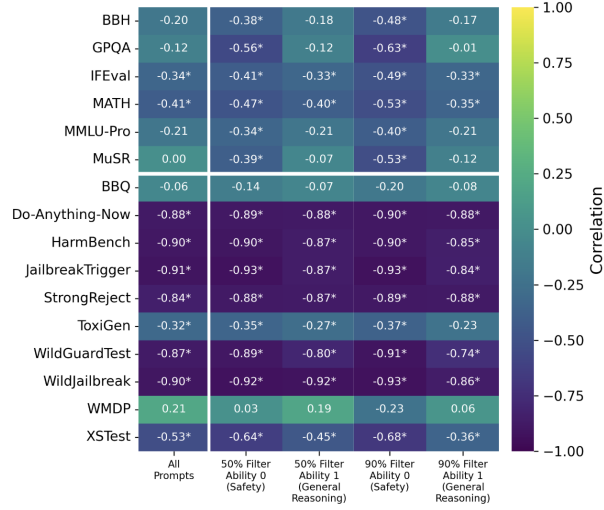


Figure 18 Correlation of each set of filtered-by-benchmark items to the MIRT ability in Dimension 0 (Safety). Asterisks indicate a p-value < 0.01.

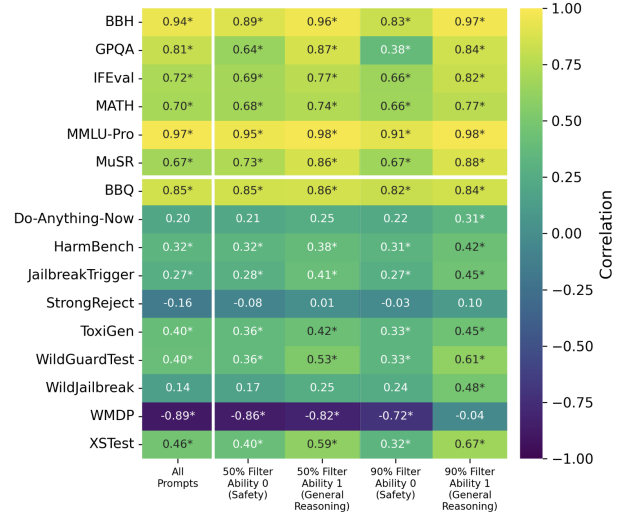


Figure 19 Correlation of each set of filtered-by-benchmark items to the MIRT ability in Dimension 1 (General Reasoning). Asterisks indicate a p-value < 0.01.

mension 1 (general reasoning) to create two sets of filtered items. In this case, we bin items inside each benchmark by their difficulty and then drop those with the lowest discrimination. As such, we keep 10% and 50% of each benchmark in addition to 10% and 50% of the total set. Figures 18, 19, and 17 show the results. Filtering based on the benchmark is generally more successful at maintaining benchmark scores even when the filtering is based on the ability not associated with that dimension.

C Item level safety classifications

We examined several trends inside the item-level safety classifications. In addition to the Safe/Unsafe/Controversial rating, Qwen3 Guard also returns the category violated. Figures 20 and 21 show the full breakdown by item discrimination bin.

We also examine safety category violations by benchmark in Figures 22 and 23. While these results largely conform to expectations and map to the benchmarks' intended ability measurements, there are some exceptions. Notably, WMDP and BBQ items are overwhelmingly rated as Safe, demonstrating the limits of this style of analysis.

Finally, we examine item classification and categorization by difficulty in both dimensions. We find that Unsafe prompts are split between high and low safety difficulties in Dimension 0, while in Dimension 1 there is a more even split with more safety items at lower difficulties.

D Ablation without low variance filter

In order to investigate the impact of filtering out items with lower than 5% or higher than 95% accuracy, we fit BENCHMIRT to all of the instances. The results are very similar to the final results presented above. Figure 30 shows the correlations between each ability dimension and average benchmark score – as found in §3.1.1, there is one dimension with a strong correlation to many of the safety benchmarks (Dimension 1 and positive in this case) and one dimension with a strong correlation to the general reasoning benchmarks (Dimension 0, also positive). This analysis also finds the misalignment of BBQ and WMDP as discussed earlier. Table D contains the predictive results on the full set of items, run with the same set up as in §3.1.5.

It again finds the same pattern as the filtered set of items, with BENCHMIRT generally out-performing all other methods on MAE and accuracy, while the per-benchmark average generally scores the best on the

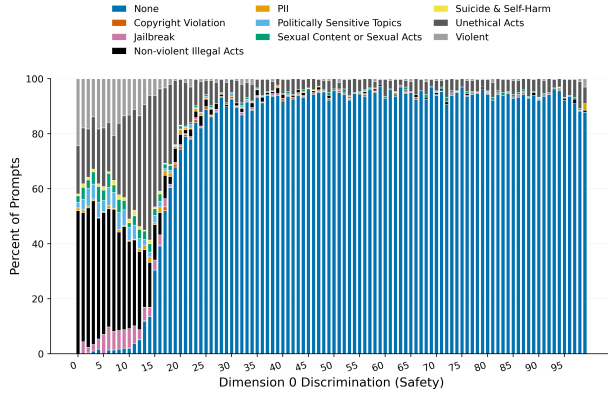


Figure 20 Unsafe and Controversial items by safety category, graphed according to item discrimination in Dimension 0 (Safety)

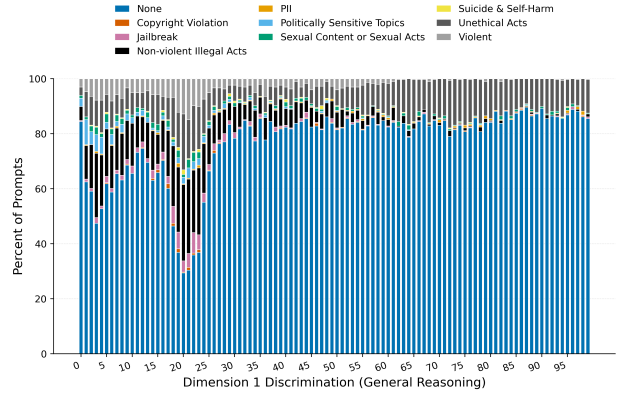


Figure 21 Unsafe and Controversial items by safety category, graphed according to item discrimination in Dimension 1 (General Reasoning)

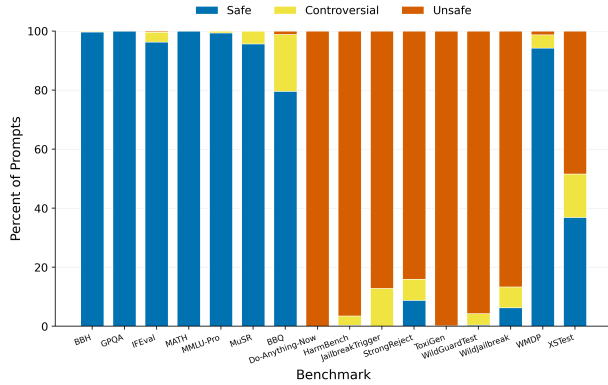


Figure 22 Unsafe and Controversial items by benchmark

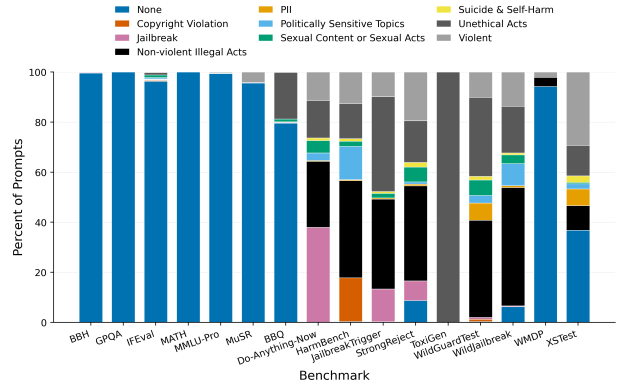


Figure 23 Unsafe and Controversial items per benchmark by safety category

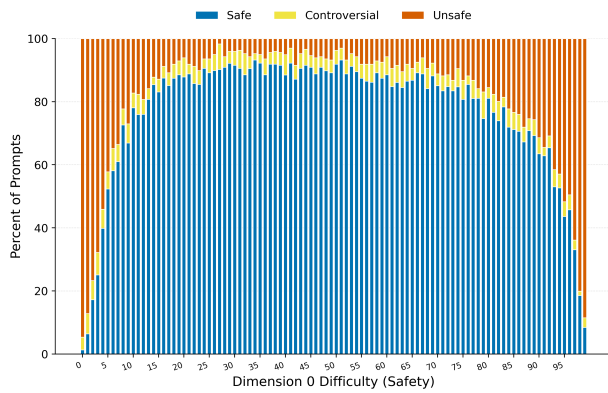


Figure 24 Unsafe and Controversial items by difficulty bin on Dimension 0 (Safety)

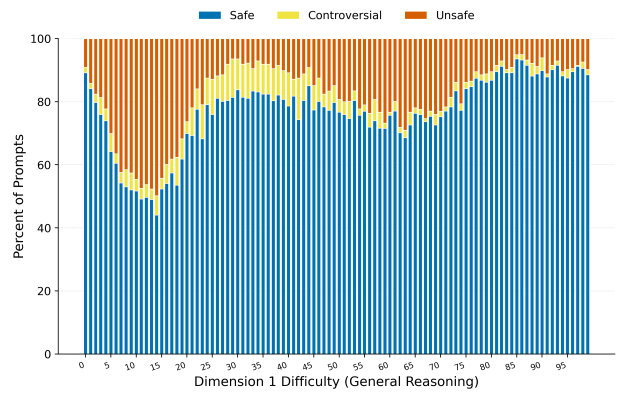


Figure 25 Unsafe and Controversial items by difficulty bin on Dimension 1 (General Reasoning)

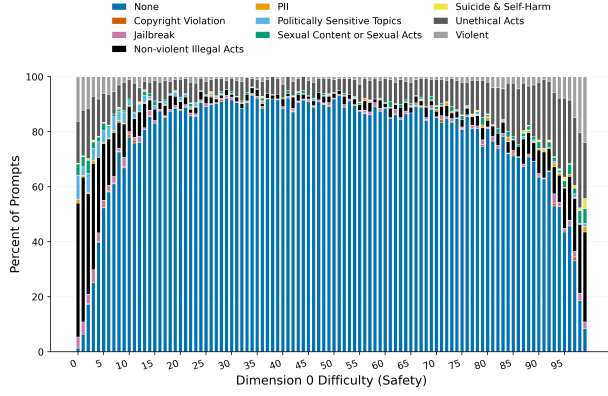


Figure 26 The safety categories of Unsafe and Controversial items by difficulty bin on Dimension 0 (Safety)

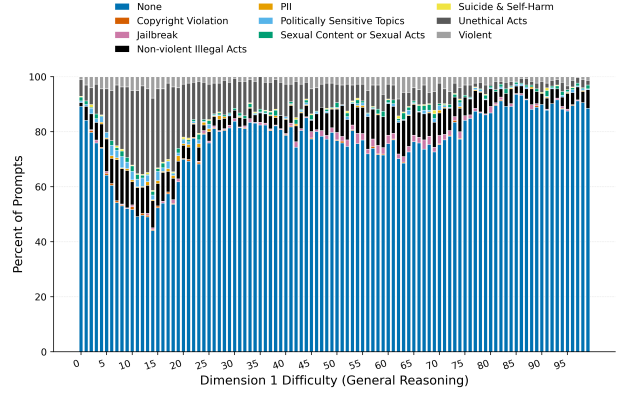


Figure 27 The safety categories of Unsafe and Controversial items by difficulty bin on Dimension 1 (General Reasoning)

	BENCHMIRT			Single-Dim IRT			Category IRT			Overall Avg			Cat. Avg			Bench. Avg			
	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	MAE	Acc	Rank	
Capability	BBH	.33.00	.75.00	.93.01	.44.00	.51.00	.57.02	.43.00	.51.00	.71.03	.48.00	.59.00	.89.00	.47.00	.55.00	.95.00	.47.00	.60.00	.98.00
	GPQA	.36.00	.74.01	.49.09	.38.00	.70.01	.37.10	.38.00	.70.01	.44.08	.49.00	.55.01	.42.07	.44.00	.66.01	.45.07	.42.00	.70.01	.39.00
	IFEval	.32.00	.75.01	.69.04	.41.00	.55.01	.45.07	.41.00	.55.01	.46.05	.47.00	.61.01	.67.04	.45.00	.62.01	.65.04	.42.00	.67.01	.93.01
	Math	.16.00	.89.00	.73.02	.20.00	.86.00	.65.03	.20.00	.86.00	.76.03	.47.00	.62.00	.74.01	.38.00	.82.00	.73.02	.21.00	.86.00	.96.01
	MMLU-Pro	.25.00	.82.00	.98.00	.34.00	.67.00	.63.02	.34.00	.67.00	.77.02	.47.00	.62.00	.93.01	.43.00	.68.00	.98.00	.41.00	.68.00	.99.00
	MUSR	.27.00	.80.00	.64.06	.37.00	.59.01	.40.09	.37.00	.59.01	.48.08	.49.00	.53.01	.36.08	.47.00	.58.01	.41.07	.48.00	.59.01	.49.09
Safety	BBQ	.25.00	.82.00	.87.01	.49.00	.27.00	.41.06	.30.00	.73.00	.79.01	.47.00	.59.00	.78.01	.39.00	.77.00	.60.02	.30.00	.79.00	.98.00
	Do Anything_Now	.29.01	.80.01	.85.01	.43.00	.60.01	.60.03	.49.00	.40.01	.26.07	.46.00	.68.01	.44.02	.48.00	.46.01	.65.02	.27.01	.81.01	.95.01
	Harmbench	.27.01	.80.01	.89.01	.46.00	.43.01	.53.06	.40.01	.57.01	.51.06	.47.00	.61.01	.56.03	.44.00	.60.01	.75.03	.39.00	.70.01	.93.01
	JailbreakTrigger	.31.01	.77.01	.90.01	.46.00	.50.01	.57.02	.46.00	.50.01	.50.04	.47.00	.65.01	.55.02	.45.00	.55.01	.77.03	.37.00	.73.01	.96.00
	StrongReject	.18.00	.87.00	.86.01	.50.00	.21.00	.40.07	.28.00	.79.00	.14.03	.51.00	.49.00	.14.02	.40.00	.73.01	.39.03	.30.00	.79.00	.98.00
	Toxigen	.12.01	.92.00	.68.03	.50.00	.08.00	.36.02	.13.00	.92.00	.76.01	.50.00	.48.00	.62.01	.35.00	.89.01	.75.01	.08.00	.94.00	.92.01
	WildGuardTest	.21.00	.85.00	.93.01	.48.00	.29.01	.53.08	.30.00	.72.01	.65.04	.48.00	.57.01	.66.02	.40.00	.72.01	.84.01	.35.00	.73.01	.96.01
	WildJailbreak	.16.01	.90.00	.94.02	.29.00	.73.00	.69.03	.46.00	.27.00	.34.06	.47.00	.62.00	.48.02	.54.00	.34.01	.69.02	.33.00	.77.00	.92.02
	WMDP	.31.00	.76.00	.84.02	.44.00	.51.01	-.36.07	.44.00	.50.01	-.54.05	.52.00	.43.00	-.76.03	.52.00	.46.01	-.59.04	.47.00	.59.00	.84.03
	XSTest	.20.01	.85.01	.71.02	.50.00	.19.01	.31.12	.25.01	.81.01	.69.04	.50.00	.51.01	.56.04	.39.00	.78.01	.66.04	.28.00	.81.01	.84.03
	micro-average	.25.00	.82.00	.99.00	.40.00	.52.00	.76.01	.34.00	.64.00	.93.01	.48.00	.59.00	.99.00	.43.00	.65.00	.99.00	.37.00	.71.00	1.0.00
	macro-average	.25.00	.82.00	.81.01	.42.00	.48.00	.44.02	.35.00	.63.00	.48.01	.48.00	.57.00	.50.01	.44.00	.64.00	.60.01	.35.00	.74.00	.88.01

Table 2 MIRT model predictions on new (model, item) pairs. Best score for each benchmark on each accuracy metric is bolded. We report the average and standard deviation (subscript) across all ten folds.

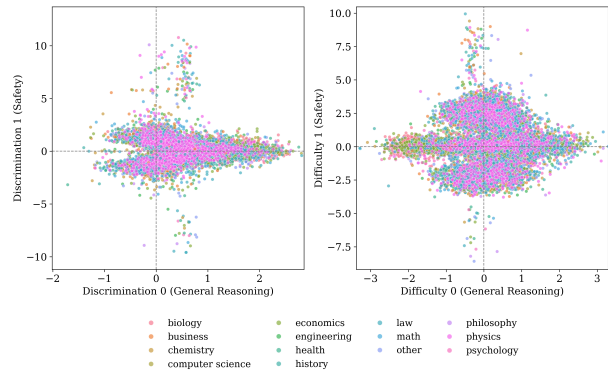


Figure 28 Item difficulty and discrimination in both dimensions for MMLU-Pro where no variance filter was applied.

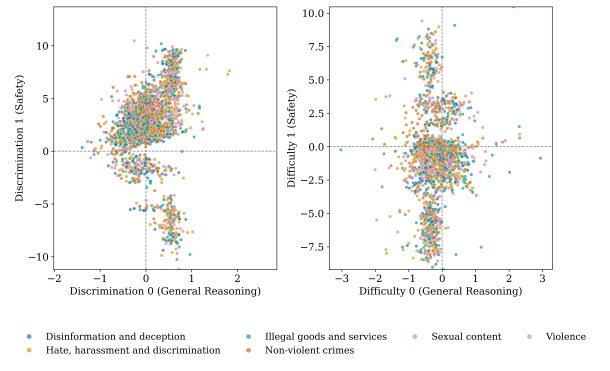


Figure 29 Item difficulty and discrimination in both dimensions for StrongReject where no variance filter was applied.)

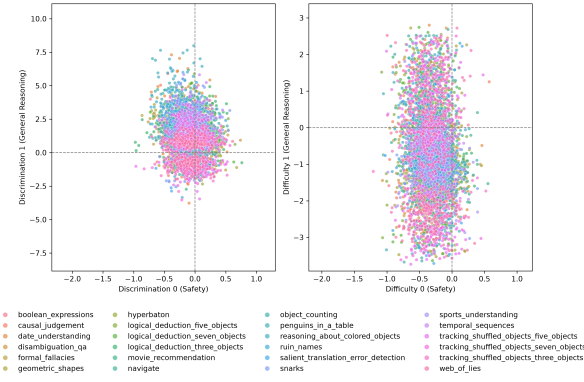


Figure 31 Item difficulty and discrimination in both dimensions for BBH.

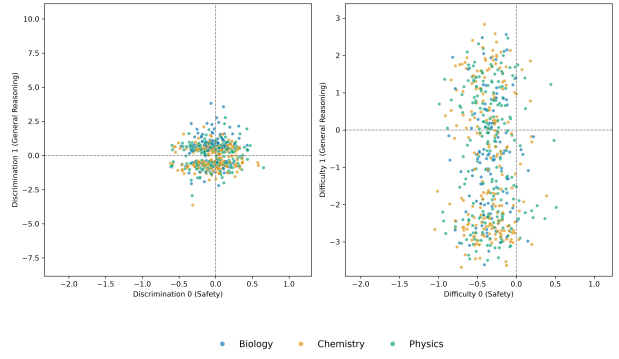


Figure 32 Item difficulty and discrimination in both dimensions for GPQA.

ranking metric.

This analysis differs from the earlier presented when auditing benchmarks directly. As Figures 28 and 29 show, there is a set of items that end up with extremely high or extremely low item parameters along the safety dimension. These items turn out to universally be the low variance items previously filtered out. All of the general reasoning benchmarks and many of the safety ones have these artifacts. The lack of signal from the model responses meant that the item parameters never got correctly updated. These items obscure the existing genuine subset patterns and interfere with item filtering without adding any information to the analysis, supporting the decision to remove them at the beginning of the pipeline.

E Qualitative analysis of benchmark make-up

Item difficulty and discrimination offer new insights into benchmark construction. Here we show the results of the same analysis discussed in §3.2 on the remaining benchmarks.

Figure 31 shows the item parameters for BBH, which is strongly aligned with the axis measuring general reasoning as expected, and well distributed across difficulty.

Figure 32 shows the item parameters for GPQA, which has a set of well-distributed difficulties across the general reasoning axis, though the item discriminations are clustered around zero. Interestingly, GPQA appears split roughly in half between positive and negatively discriminating items.

Figure 33 has the item difficulties and discriminations for MuSR. Like MMLU-Pro, the parameters are primarily clustered around the general reasoning dimension, indicating a strong alignment to the stated ability, with generally positive item discrimination scores.

As seen in Figure 34, the MATH benchmark appears generally similar to the other general reasoning benchmarks, though the majority of its items have higher difficulty scores.

Figure 35 shows that Do-Anything-Now is well aligned to and

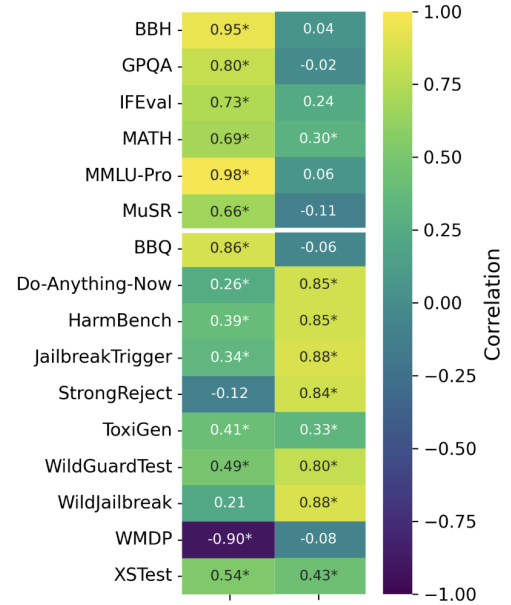


Figure 30 The correlation between ability and LLM average on each unfiltered benchmark. Starred numbers denote p-values < 0.01.

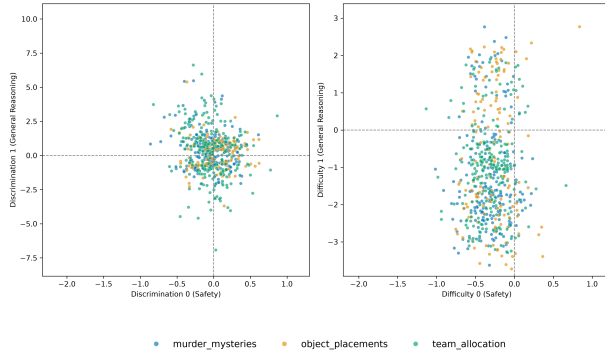


Figure 33 Item difficulty and discrimination in both dimensions for MuSR.

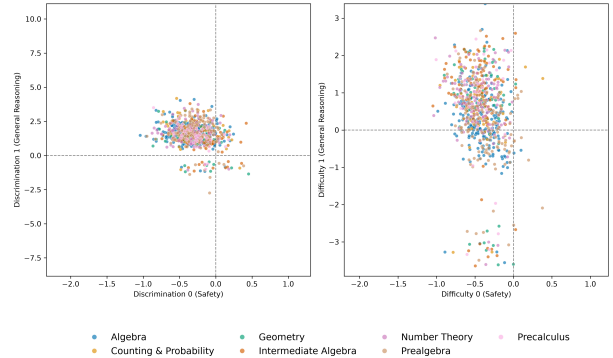


Figure 34 Item difficulty and discrimination in both dimensions for MATH.

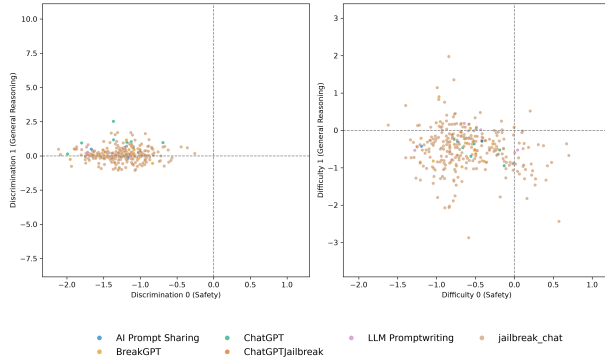


Figure 35 Item difficulty and discrimination in both dimensions for Do-Anything-Now.

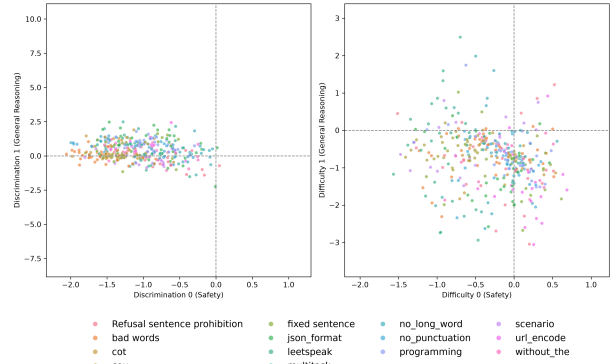


Figure 36 Item difficulty and discrimination in both dimensions for TrustLLM-JailbreakTrigger.

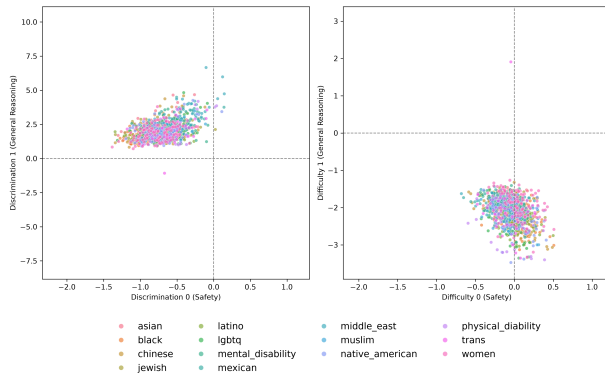


Figure 37 Item difficulty and discrimination in both dimensions for ToxiGen.

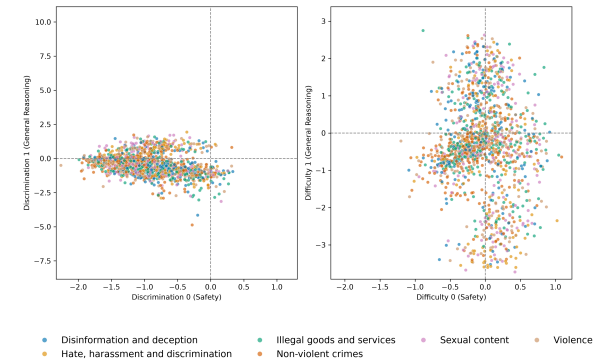


Figure 38 Item difficulty and discrimination in both dimensions for StrongReject.

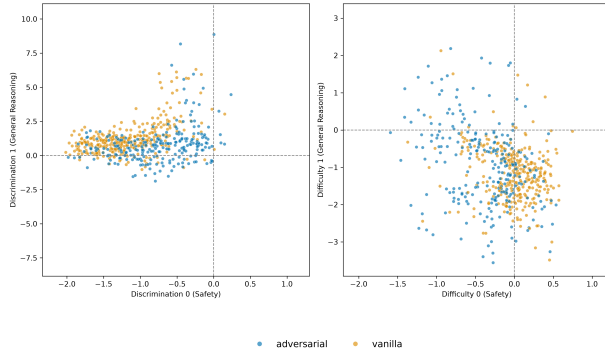


Figure 39 Item difficulty and discrimination in both dimensions for WildGuardTest.

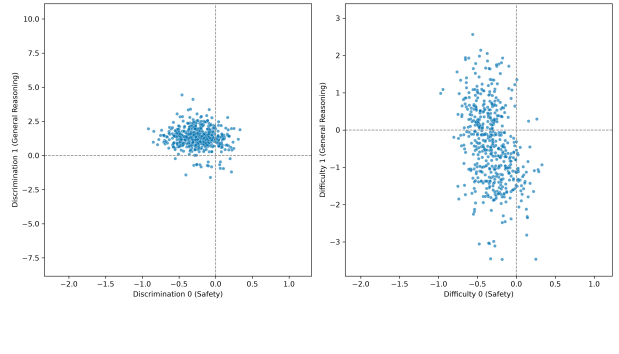


Figure 40 Item difficulty and discrimination in both dimensions for IFEval.

discriminative of safety, though the items are in general not difficult. This may reflect Do-Anything-Now’s narrow, comparatively older jailbreak style.

Very similarly to Do-Anything-Now, Figure 36 shows that TrustLLM-JailbreakTrigger is well-aligned to safety, though not a difficult benchmark. TrustLLM-JailbreakTrigger does also demonstrate some separation between how discriminative different subsets of the benchmark are, though not as distinctive as HarmBench it is clear that several of the most discriminative items are from the "bad words" subset.

Figure 37 has the item parameters for ToxiGen. Interestingly, ToxiGen discrimination scores indicate it is measuring both dimensions, while the difficulty scores are strongly negative on the general reasoning axis. This pattern is likely a reflection of the general saturation of ToxiGen.

From Figure 38, we can see that StrongReject item discrimination scores are tightly clustered around the safety axis, though the item difficulty scores show a greater spread along the general reasoning axis. This may be a reflection of the puzzle-like nature of some of StrongReject’s jailbreaks, which are more difficult to understand.

Figure 39 shows a separation in the subsets of WildGuardTest, where vanilla prompts are more consistently discriminative than adversarial ones, though the adversarial ones may be more difficult. The general negative difficulty scores likely reflects a higher saturation of the benchmark score – though not as pronounced as ToxiGen, many models incorrectly answering questions may have generally lower abilities.

As seen in Figure 40, the IFEval items are almost all positively discriminative, indicating a well-made benchmark, and the difficulties have a large spread on the general reasoning axis while remaining more closely centered on the safety axis.

F Additional dimensions

Further results on the three and four dimensional models discussed in §3.1.7 are presented here. Table F shows the accuracy results on each benchmark .

Figure 41 shows the correlations between each dimension in the different models.

G LLMs included in the model

From the LLM models available on the OpenLLM Leaderboard v2, which include models open-sourced up to March of 2025, we filtered to those marked as from official providers, currently available on Hugging Face, not flagged as possibly contaminated, and not classified as 'pretrained', 'base merges and moerges', or

Table 3 Model predictions on new (model, item) pairs for the three and four dimensional MIRT models. We report the average and standard deviation (subscript) across all ten folds.

		Three-Dimensional MIRT			Four-Dimensional MIRT		
		MAE	Acc	Rank	MAE	Acc	Rank
Capability	BBH	.33 _{.00}	.75 _{.00}	.93 _{.00}	.32 _{.00}	.76 _{.00}	.95 _{.00}
	GPQA	.36 _{.00}	.73 _{.01}	.50 _{.08}	.35 _{.00}	.73 _{.01}	.52 _{.07}
	IFEval	.33 _{.01}	.74 _{.01}	.70 _{.04}	.33 _{.00}	.75 _{.01}	.73 _{.04}
	Math	.24 _{.00}	.83 _{.00}	.78 _{.02}	.24 _{.00}	.83 _{.00}	.78 _{.01}
	MMLU-Pro	.28 _{.00}	.80 _{.00}	.97 _{.00}	.27 _{.00}	.80 _{.00}	.98 _{.00}
	MUSR	.32 _{.00}	.76 _{.00}	.60 _{.06}	.32 _{.00}	.76 _{.00}	.59 _{.06}
Safety	BBQ	.24 _{.00}	.82 _{.00}	.88 _{.01}	.24 _{.00}	.82 _{.00}	.89 _{.01}
	Do_Anything_Now	.29 _{.01}	.80 _{.01}	.86 _{.01}	.28 _{.01}	.80 _{.01}	.86 _{.01}
	Harmbench	.29 _{.01}	.79 _{.00}	.89 _{.01}	.28 _{.01}	.79 _{.01}	.90 _{.01}
	JailbreakTrigger	.31 _{.01}	.77 _{.01}	.91 _{.01}	.30 _{.00}	.77 _{.01}	.91 _{.01}
	StrongReject	.28 _{.00}	.80 _{.00}	.91 _{.01}	.28 _{.00}	.80 _{.00}	.92 _{.01}
	Toxigen	.13 _{.00}	.91 _{.00}	.69 _{.01}	.12 _{.01}	.91 _{.00}	.76 _{.01}
	WildGuardTest	.26 _{.00}	.81 _{.00}	.93 _{.01}	.25 _{.00}	.81 _{.00}	.94 _{.01}
	WildJailbreak	.18 _{.01}	.88 _{.00}	.93 _{.01}	.18 _{.01}	.88 _{.00}	.93 _{.01}
	WMDP	.31 _{.00}	.76 _{.00}	.83 _{.02}	.31 _{.00}	.76 _{.00}	.83 _{.02}
	XSTest	.23 _{.00}	.83 _{.01}	.75 _{.03}	.23 _{.01}	.83 _{.01}	.76 _{.03}
micro-average		.27 _{.00}	.80 _{.00}	.99 _{.00}	.27 _{.00}	.80 _{.00}	1.0 _{.00}
macro-average		.27 _{.00}	.80 _{.00}	.82 _{.00}	.27 _{.00}	.80 _{.00}	.83 _{.01}

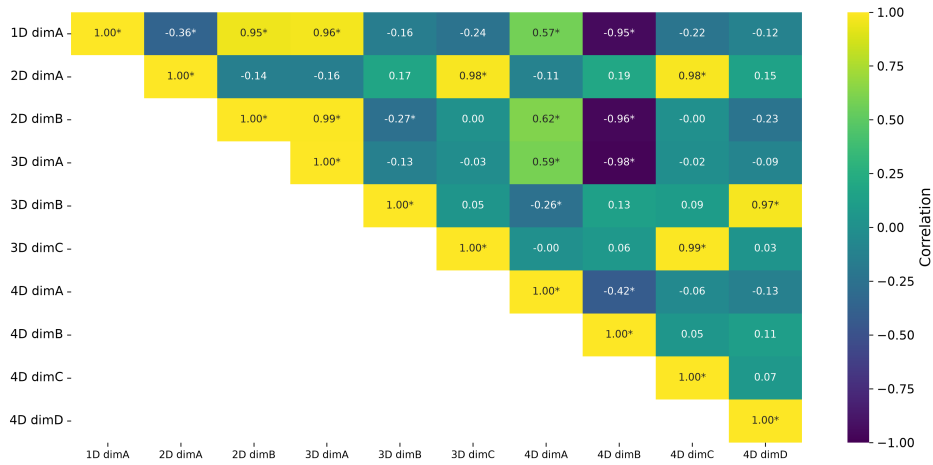


Figure 41 The Pearson Correlation Coefficient between each ability in the (M)IRT models. Starred numbers denote p-values < 0.01. Dimensions for each model are referred to by ABCD to differentiate them.

Model	License
01-ai/Yi-1.5-34B-Chat-16K (Yi-, 2024)	apache-2.0
01-ai/Yi-34B-Chat (Yi-, 2023)	apache-2.0
abacusai/Liberated-Qwen1.5-14B(Lib, 2024)	Tongyi Qian
abacusai/Llama-3-Smaug-8B(Lla, 2024c)	Llama2
abacusai/Smaug-72B-v0.1(Sma, 2024a)	Tongyi Qian
abacusai/Smaug-Qwen2-72B-Instruct (Sma, 2024b)	Tongyi Qian
allenai/Llama-3.1-Tulu-3-8B(Lla, 2024h)	Llama3.1
allenai/OLMo-2-1124-7B-Instruct(OLM, 2024a)	apache-2.0
allenai/OLMoE-1B-7B-0924-Instruct(OLM, 2024b)	apache-2.0
allenai/Llama-3.1-Tulu-3-70B(Lla, 2024g)	Llama3.1
amd/AMD-Llama-135m(AMD, 2024)	apache-2.0
berkeley-nest/Starling-LM-7B-alpha(Sta, 2023)	apache-2.0
CausalLM/14B(14B, 2023)	WTFPL
cognitivecomputations/dolphin-2.9.2-Phi-3-Medium-abliterated(dol, 2024c)	MIT
cognitivecomputations/Dolphin3.0-R1-Mistral-24B(Dol, 2025)	unknown
cognitivecomputations/dolphin-2.9.1-llama-3-70b(dol, 2024a)	Llama3
cognitivecomputations/dolphin-2.9.1-yi-1.5-34b(dol, 2024b)	apache-2.0
cognitivecomputations/dolphin-2.9.2-qwen2-72b(dol, 2024d)	Tongyi Qian
cognitivecomputations/dolphin-2.9.3-Yi-1.5-34B-32k(dol, 2024e)	apache-2.0
cognitivecomputations/Dolphin3.0-Llama3.1-8B(Dol, 2024a)	Llama3.1
cognitivecomputations/Dolphin3.0-Llama3.2-1B(Dol, 2024b)	Llama3.2
cognitivecomputations/Dolphin3.0-Qwen2.5-0.5B(Dol, 2024c)	apache-2.0
CohereForAI/c4ai-command-r-plus-08-2024(c4a, 2024)	CC BY-NC 4.0
deepseek-ai/deepseek-llm-7b-chat(dee, 2023)	deepseek
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B(Dee, 2025e)	MIT
deepseek-ai/DeepSeek-R1-Distill-Llama-8B(Dee, 2025b)	MIT
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B(Dee, 2025d)	MIT
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B(Dee, 2025c)	MIT
google/gemma-1.1-7b-it(gem, 2024a)	gemma
google/gemma-2-27b-it(gem, 2024b)	gemma
gradientai/Llama-3-8B-Instruct-Gradient-1048k(Lla, 2024a)	Llama3
HuggingFaceH4/zephyr-7b-beta(zep, 2023)	MIT
HuggingFaceTB/SmolLM2-360M-Instruct(Smo, 2024)	apache-2.0
ibm-granite/granite-3.0-8b-instruct(gra, 2024)	apache-2.0
ibm-granite/granite-3.2-8b-instruct(gra, 2025)	apache-2.0
ibm/merlinit-7b(mer, 2024)	apache-2.0
Intel/neural-chat-7b-v3-3(neu, 2023)	apache-2.0
internlm/internlm2_5-7b-chat(int, 2024)	apache-2.0
LLM360/K2-Chat(K2-, 2024)	apache-2.0

Model	License
lmsys/vicuna-7b-v1.5(vic, 2023)	Llama2
meta-llama/Llama-3.1-70B-Instruct(Lla, 2024d)	Llama3.1
meta-llama/Llama-3.1-8B-Instruct(Lla, 2024e)	Llama3.1
meta-llama/Llama-3.2-1B-Instruct(Lla, 2024i)	Llama3.2
meta-llama/Llama-3.2-3B-Instruct(Lla, 2024j)	Llama3.2
meta-llama/Llama-3.3-70B-Instruct(Lla, 2024k)	Llama3.3
meta-llama/Meta-Llama-3-8B-Instruct(Met, 2024a)	Llama3
microsoft/Phi-3-medium-4k-instruct(Phi, 2024a)	MIT
microsoft/Phi-3-mini-4k-instruct(Phi, 2024b)	MIT
microsoft/Phi-3.5-MoE-instruct(Phi, 2024c)	MIT
microsoft/phi-4 (phi, 2024)	MIT
mistralai/Minstral-8B-Instruct-2410 (Min, 2024)	Mistral Research License
mistralai/Mistral-7B-Instruct-v0.2 (Mis, 2023a)	apache-2.0
mistralai/Mistral-Large-Instruct-2411 (Mis, 2024a)	Mistral Research License
mistralai/Mistral-Nemo-Instruct-2407 (Mis, 2024b)	apache-2.0
mistralai/Mistral-Small-Instruct-2409 (Mis, 2024c)	Mistral Research License
mlabonne/AlphaMonarch-7B (Alp, 2024)	CC BY-NC 4.0
mlabonne/Daredevil-8B-abliterated (Dar, 2024)	Llama3
mlabonne/Hermes-3-Llama-3.1-70B-lorabliterated (Her, 2024c)	Llama3.1
mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated (Met, 2024b)	Llama3.1
mlabonne/NeuralDaredevil-8B-abliterated (Neu, 2024)	Llama3
Nexusflow/NexusRaven-V2-13B (Nex, 2023)	nexusflow.ai
NousResearch/DeepHermes-3-Mistral-24B-Preview (Dee, 2025a)	apache-2.0
NousResearch/Hermes-2-Pro-Mistral-7B (Her, 2024a)	apache-2.0
NousResearch/Hermes-3-Llama-3.1-70B (Her, 2024b)	Llama3
NousResearch/Nous-Hermes-2-SOLAR-10.7B (SOL, 2023)	apache-2.0
NousResearch/Nous-Hermes-llama-2-7b (Nou, 2023)	MIT
nvidia/AceInstruct-1.5B (Ace, 2025a)	CC BY-NC 4.0
nvidia/AceInstruct-72B (Ace, 2025b)	CC BY-NC 4.0
nvidia/AceMath-72B-Instruct (Ace, 2025c)	CC BY-NC 4.0
nvidia/AceMath-7B-Instruct (Ace, 2025d)	CC BY-NC 4.0
nvidia/Llama-3.1-Nemotron-70B-Instruct-HF (Lla, 2024f)	Llama3.1
nvidia/OpenMath2-Llama3.1-8B (Ope, 2024)	Llama3.1
Open-Orca/Mistral-7B-OpenOrca (Mis, 2023b)	apache-2.0
OpenAssistant/oasst-sft-1-pythia-12b (oas, 2023)	apache-2.0
openchat/openchat-3.5-0106 (ope, 2024)	apache-2.0
PrimeIntellect/INTELLECT-1-Instruct(INT, 2024)	apache-2.0
PygmalionAI/pygmalion-6b (pyg, 2023)	creativeml-openrail

Model	License
Qwen/Qwen2-72B-Instruct (Qwe, 2024a)	Tongyi Qianwen
Qwen/Qwen2-VL-7B-Instruct (Qwe, 2024b)	apache-2.0
Qwen/Qwen2.5-0.5B-Instruct (Qwe, 2024c)	apache-2.0
Qwen/Qwen2.5-1.5B-Instruct (Qwe, 2024d)	apache-2.0
Qwen/Qwen2.5-14B-Instruct (Qwe, 2024e)	apache-2.0
Qwen/Qwen2.5-72B-Instruct (Qwe, 2024f)	Qwen License
Qwen/Qwen2.5-Coder-7B-Instruct (Qwe, 2024g)	apache-2.0
Qwen/QwQ-32B (QwQ, 2025)	apache-2.0
refuelai/Llama-3-Refueled (LLa, 2024b)	apache-2.0
rhymes-ai/Aria (Ari, 2024)	apache-2.0
speakeash/Bielik-11B-v2.3-Instruct (Bie, 2024)	apache-2.0
stabilityai/stablelm-2-zephyr-1_6b (sta, 2024)	Stability AI
teknium/OpenHermes-13B (Ope, 2023a)	MIT
teknium/OpenHermes-2.5-Mistral-7B (Ope, 2023b)	apache-2.0
TencentARC/LLaMA-Pro-8B (LLa, 2024a)	Llama2
TencentARC/LLaMA-Pro-8B-Instruct (LLa, 2024b)	Llama2
tiiuae/Falcon3-7B-Instruct (Fal, 2024)	Falcon LLM License
tiiuae/Falcon-7B-Instruct (Fal, 2023)	apache-2.0
TinyLlama/TinyLlama-1.1B-Chat-v1.0(Tin, 2023)	apache-2.0
togethercomputer/GPT-NeoXT-Chat-Base-20B(GPT, 2023)	apache-2.0
upstage/SOLAR-10.7B-Instruct-v1.0(SOL, 2023)	CC-BY-NC-4.0
WizardLMTeam/WizardLM-13B-V1.2 (Wiz, 2023a)	Llama2
WizardLMTeam/WizardLM-70B-V1.0(Wiz, 2023b)	Llama2

'continuously pretrained'.⁷ Given the approximately 250 LLMs remaining, we selected 100 models to maximize coverage across ability levels, model sizes, model families, reasoning (trained to produce a Chain of Thought) vs instruct models, and estimated safety ability. Tables G and G gives more information for each model.

⁷meta-llama/Llama-3.1-8B-Instruct is included, though the leaderboard has it classified as a pretrained model and meta-llama/Llama-3.1-8B as the finetuned.