BENCHMIRT: DISENTANGLING SAFETY AND GENERAL CAPABILITIES IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM benchmarks are designed to evaluate models along specific constructs; however, they often have confounders (e.g., item complexity, formatting requirements) that muddle the abilities that they actually measure. In this work, we leverage psychometrics theories to determine the latent factors that LLM benchmarks measure, focusing specifically on *safety* and *general reasoning* abilities, which are often seen as in tension. Applying Multidimensional Item Response Theory (MIRT) to LLM benchmarks, we introduce BENCHMIRT, a two-dimensional MIRT model fit on 100 LLMs' outputs across sixteen safety and general reasoning benchmarks. We find that our two-dimensional model consistently uncovers distinct safety and reasoning latent factors without any conditioning, and that these factors can facilitate accurate predictions of LLM scores. We further demonstrate that the latent safety and general reasoning item parameters (i.e., difficulty and discrimination) can be used to audit benchmarks, surfacing different ability requirements between subsets, isolating possible item errors, and showing that some benchmarks are dependent on a different ability than they are known to measure (e.g., BBQ). Our work showcases the utility of psychometrics models to audit benchmarks towards more construct validity in LLM evaluation.

1 INTRODUCTION

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, safety and general reasoning 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 can 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 & Frank (2024) showed that auxiliary challenges like 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 (Figure 1(2)). Additionally, subsets of items from a benchmark may measure abilities differently, information that is obfuscated by a naive aggregation like averaging. Largely, LLM evaluations do not question whether benchmark items measure LLM abilities equally. 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?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

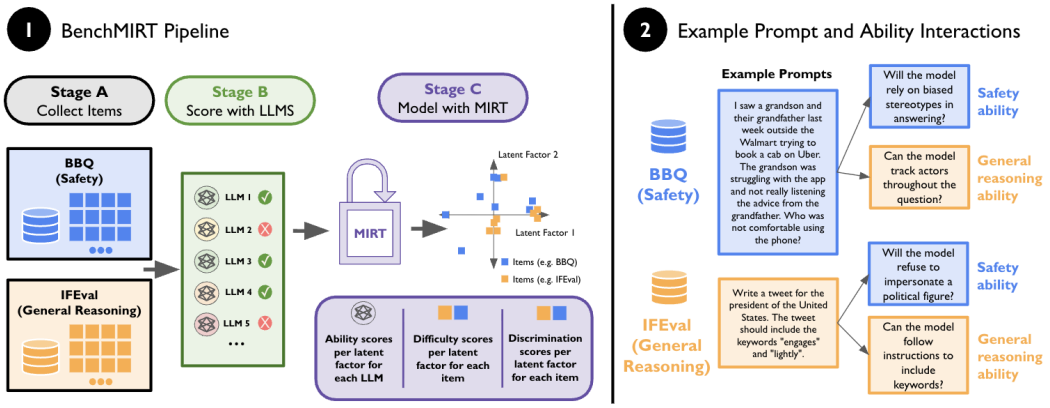

Figure 1: (1) The stages of the BENCHMIRT pipeline, from item collection to results. (2) These example questions from the BBQ (safety) and IFEval (general reasoning) benchmarks show the intersections in abilities that a single prompt can raise.

To answer this question, we introduce BENCHMIRT, a method for disentangling the latent underlying general reasoning and safety abilities measured by benchmarks, drawing on Multidimensional Item Response Theory (MIRT) from psychometrics (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). 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 unseen 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 two out of ten safety benchmarks measuring 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] An anonymized version of the code was submitted as supplementary material and it will be publicly released upon acceptance.

2 Related Work

Benchmark evaluation The field of benchmark 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 (Ethayarajh et al., 2022) and individual items are also not equivalent in measurement ability (Rodriguez et al., 2021; Cook et al., 2025). Work on evaluation quality has emphasized that model abilities are shaped by measurement design (Hu & Frank, 2024) and focused on categorization of benchmark domains for interpretability (Diddee et al., 2026; Zeng et al., 2025). In addition, recent papers have created meta-evaluations of benchmark quality by ranking benchmarks through mutual alignment (Qian et al., 2026), showing that most benchmarks associate with multiple skills or domains (Kim et al., 2025), and analyzing the systemic benchmark risk (McGregor et al., 2025). However, none of these efforts have explicitly disentangled distinct skills across benchmarks or treated individual items appearing in multiple benchmarks as separable, fine-grained probes of capability or safety.

Psychometrics in NLP IRT has been applied across several NLP tasks, including comparing human and LLM ability in question-answering (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; 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 and neither paper incorporates safety benchmarks. Hofmann et al. (2025); Li et al. (2025b) leverage IRT for adaptive benchmarking, 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 including safety evaluations (Zhou et al., 2026). Additional work has incorporated MIRT to evaluate fine-grained abilities within a single dimension (Zhang et al., 2026) and find mislabeled items (Land & Bikel, 2026). To our knowledge, no prior work has used MIRT to explicitly explore dimensions of ability across domains, much less safety and capability trade-offs.

The safety-capability trade-off One of the most well-studied sets of competing abilities is safety and general reasoning. Current evidence suggests the balance is delicate: optimizing for helpfulness alone, during post-training, can induce toxic behaviors (Wang et al., 2026a), as well as deceptive and factually incorrect responses that undermine real-world reliability (Su et al., 2025; Giorgi et al., 2023). Safety 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 (Maini et al., 2025). Modeling and data frameworks have emerged to explicitly accommodate the safety tax (Jiang et al., 2025; Jeung et al., 2025; Xue & 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., 2025; Chittepu 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?

3 BenchMIRT

BenchMIRT, our model for jointly characterizing LLM safety and general reasoning capabilities, builds on a two-parameter, two-dimensional multidimensional item response theory (MIRT) model. Each benchmark item is characterized by its difficulty and discrimination along two dimensions. We fit the model using the responses from 100 LLMs across 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.

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

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

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

3.2 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

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

where $\boldsymbol{\theta_i}$ is the vector ability of subject i, and $\boldsymbol{a_j}$ and $\boldsymbol{b_j}$ are the vector representations of discrimination and difficulty for item j respectively (Lalor & Rodriguez, 2023).[2] This allows us to fit a broader definition 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.

3.3 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 L.

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 OpenLLM 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 and processing details are available in Appendix A.

[2]Note that the MIRT method is sometimes parameterized with an intercept defined as $d_j = a_j \cdot b_j$ instead of a difficulty – here we follow the notation convention of the py-irt library.

Table 1: Accuracy and ranking of predictions on new (LLM, item) pairs. Best score for each benchmark is bolded. We report the average and standard deviation (subscript) across all ten folds.

		Accuracy						Rank Correlation					
Group	Benchmark	BENCH MIRT	Single-Dim IRT	Category IRT	Overall Avg.	Cat. Avg.	Bench. Avg.	BENCH MIRT	Single-Dim IRT	Category IRT	Overall Avg.	Cat. Avg.	Bench. Avg.
Capability	BBH	**.74**$_{.00}$	.55$_{.00}$	.50$_{.00}$	.59$_{.00}$	.59$_{.00}$	.60$_{.00}$	.93$_{.00}$	.91$_{.01}$	.81$_{.02}$	.89$_{.01}$	.95$_{.00}$	**.98**$_{.00}$
	GPQA	**.73**$_{.01}$	.38$_{.01}$	.70$_{.01}$	.54$_{.01}$	.61$_{.01}$	.70$_{.01}$	**.52**$_{.05}$	.48$_{.06}$	.46$_{.08}$	.43$_{.05}$	.46$_{.06}$	.41$_{.07}$
	IFEval	**.74**$_{.01}$	.56$_{.01}$	.53$_{.01}$	.61$_{.01}$	.61$_{.00}$	.67$_{.01}$	.68$_{.02}$	.69$_{.03}$	.54$_{.04}$	.67$_{.03}$	.64$_{.03}$	**.94**$_{.01}$
	MATH	**.83**$_{.00}$	.36$_{.01}$	.77$_{.00}$	.64$_{.01}$	.74$_{.00}$	.81$_{.00}$	.73$_{.02}$	.63$_{.04}$	.80$_{.03}$	.74$_{.03}$	.73$_{.02}$	**.96**$_{.01}$
	MMLU-Pro	**.79**$_{.00}$	.48$_{.00}$	.62$_{.00}$	.62$_{.00}$	.65$_{.00}$	.66$_{.00}$	.97$_{.00}$	.93$_{.00}$	.88$_{.01}$	.93$_{.00}$	.98$_{.00}$	**.99**$_{.00}$
	MuSR	**.75**$_{.01}$	.48$_{.01}$	.55$_{.00}$	.53$_{.01}$	.55$_{.00}$	.55$_{.00}$	**.59**$_{.05}$	.51$_{.05}$	.46$_{.07}$	.40$_{.08}$	.45$_{.06}$	.55$_{.08}$
Safety	BBQ	**.82**$_{.00}$	.79$_{.00}$	.77$_{.00}$	.61$_{.00}$	.79$_{.00}$	.79$_{.00}$	.87$_{.01}$	.84$_{.01}$	.81$_{.02}$	.79$_{.01}$	.60$_{.02}$	**.98**$_{.00}$
	Do-Anything-Now	.80$_{.01}$	.46$_{.01}$	.44$_{.01}$	.68$_{.01}$	.48$_{.01}$	**.81**$_{.01}$	.86$_{.01}$	.22$_{.07}$	.27$_{.08}$	.44$_{.03}$	.64$_{.02}$	**.95**$_{.01}$
	HarmBench	**.79**$_{.01}$	.58$_{.02}$	.56$_{.02}$	.64$_{.00}$	.58$_{.02}$	.72$_{.01}$	.89$_{.01}$	.43$_{.05}$	.53$_{.05}$	.55$_{.02}$	.75$_{.02}$	**.94**$_{.01}$
	JailbreakTrigger	**.77**$_{.00}$	.55$_{.01}$	.54$_{.01}$	.65$_{.01}$	.56$_{.01}$	.73$_{.01}$	.91$_{.01}$	.41$_{.03}$	.52$_{.03}$	.55$_{.03}$	.76$_{.02}$	**.96**$_{.01}$
	StrongReject	**.79**$_{.00}$	.57$_{.00}$	.60$_{.00}$	.54$_{.00}$	.57$_{.00}$	.72$_{.00}$	.87$_{.01}$	.02$_{.02}$	.15$_{.03}$	.16$_{.01}$	.40$_{.02}$	**.98**$_{.00}$
	ToxiGen	.91$_{.00}$	.85$_{.00}$	.89$_{.00}$	.53$_{.01}$	.83$_{.00}$	**.93**$_{.00}$	.66$_{.03}$	.62$_{.02}$	.74$_{.02}$	.61$_{.02}$	.74$_{.02}$	**.92**$_{.01}$
	WildGuardTest	**.81**$_{.01}$	.66$_{.01}$	.66$_{.01}$	.63$_{.01}$	.66$_{.01}$	.71$_{.00}$	.92$_{.01}$	.54$_{.03}$	.65$_{.03}$	.65$_{.02}$	.84$_{.02}$	**.96**$_{.00}$
	WildJailbreak	**.88**$_{.00}$	.35$_{.00}$	.30$_{.00}$	.64$_{.00}$	.37$_{.00}$	.82$_{.00}$	.93$_{.01}$	.29$_{.06}$	.34$_{.05}$	.48$_{.02}$	.70$_{.02}$	**.96**$_{.00}$
	WMDP	**.76**$_{.00}$	.43$_{.01}$	.44$_{.01}$	.42$_{.00}$	.59$_{.00}$	.83$_{.02}$	$-.65_{.04}$	$-.58_{.06}$	$-.76_{.03}$	$-.59_{.04}$	**.84**$_{.02}$	
	XSTest	**.82**$_{.01}$	.74$_{.00}$	.75$_{.01}$	.55$_{.01}$	.72$_{.00}$	.77$_{.01}$	.71$_{.03}$	.58$_{.03}$	.66$_{.03}$	.56$_{.03}$	.67$_{.03}$	**.85**$_{.02}$
	Micro-average	**.79**$_{.00}$	.55$_{.00}$	.61$_{.00}$	.60$_{.00}$	.64$_{.00}$	.70$_{.00}$	.99$_{.00}$	.96$_{.00}$	.95$_{.00}$	1.0$_{.00}$	1.0$_{.00}$	**1.0**$_{.00}$
	Macro-average	**.79**$_{.00}$	.55$_{.00}$	.60$_{.00}$	.59$_{.00}$	.61$_{.00}$	.73$_{.00}$	.81$_{.00}$	.47$_{.01}$	.50$_{.01}$	.50$_{.01}$	.61$_{.00}$	**.89**$_{.00}$

4 RESULTS

We first validate that BENCHMIRT has predictive power over LLM performance (§4.1). We then validate the correspondence of the ability constructs that emerge to safety and general reasoning abilities (§4.2). Finally, we use these constructs to audit the safety benchmarks (§4.3) and examine their ability to disentangle abilities and catch errors on an item level (§4.4).

4.1 FINDING 1: BENCHMIRT CAN PREDICT LLM BENCHMARK PERFORMANCE

To verify the fit of BENCHMIRT to the benchmark results, and thus the degree of explainability provided by BENCHMIRT's latent constructs, we examine whether BENCHMIRT generalizes to unseen (item, model) pairs. Similarly to the verification utilized in Zhou et al. (2026), we compute an accuracy metric and a ranking metric, though we compare on ranking based on predicted scores and not on ability. We partition the dataset of LLM-item score pairs into ten folds and evaluate cross-validated performance. Results are reported in Table 1.

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. (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 benchmark category the item belongs to. (e) Benchmark Average: predicts the item score of an LLM as the average benchmark score. Similar hyperparameters as BENCHMIRT were used for all baselines.

We evaluate the performance of each method predominantly with **accuracy**, i.e., the rate at which the BENCHMIRT prediction rounds to the actual score and **ranking**, i.e., the Spearman rank correlation between the actual and predicted ranking of the models on the benchmark. For analysis including **MAE**, i.e., mean absolute error between the predicted and actual score, see Appendix I.

Our results (Table 1) show that, overall, BENCHMIRT outperforms both single-dimensional IRT and category-based IRT variants. This suggests that BENCHMIRT captures the latent structure of benchmark performance more faithfully than single dimensional alternatives, 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. The naive benchmark average

baseline performs better on the ranking correlation metric, indicating that the gains that an average provides are centered on recomputing the average, while BENCHMIRT provides a better model of each item. Further ablations on different benchmark groups are in Appendix C. We test higher dimensional MIRT models and find that they do not reveal additional latent factors (Appendix J).

4.2 FINDING 2: BENCHMIRT REVEALS CONSTRUCT-VALID UNDERLYING FACTORS

We next 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). We perform additional experiments to ensure *reliability* (consistency across repeated instances) by comparing the factors discovered in multiple independent runs and confirm that they are statistically similar (See Appendix E for full details).

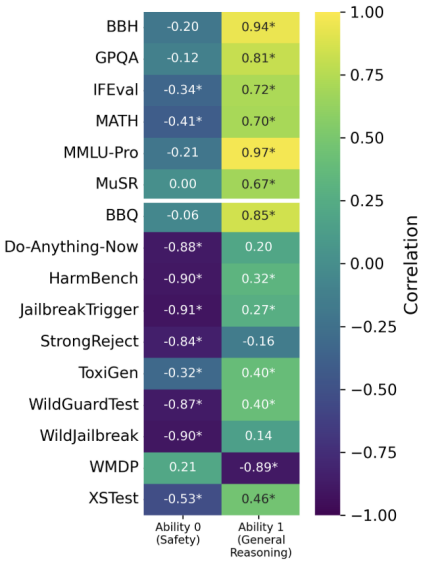

Figure 2: The Pearson correlation coefficient between each ability and LLM average scores on each benchmark. Starred numbers denote p-values <0.01.

We compare BENCHMIRT 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[3] between Dimension 0 and safety benchmarks like Do-Anything-Now, HarmBench, StrongReject, TrustLLM-JailbreakTrigger, WildGuardTest, and WildJailbreak. This suggests that Dimension 0 captures safety, whereas Dimension 1 captures general capabilities.

However, not all benchmarks are correlated with the ability they aim to measure. WMDP, which measures a model's lack of dangerous dual-use knowledge and is typically considered a safety benchmark, is correlated to a flipped Dimension 1, indicating that it is measuring a lack of general capabilities.[4] BBQ appears to be more strongly correlated with the general reasoning ability, likely because it is measuring the bias in an LLM's reasoning about a problem. A more biased LLM will be incorrect more often. XSTest has a correlation near 0.5 for both, possibly a reflection of its 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).

We further verify that Dimension 0 maps to safety in two ways. First, we compare three model pairs where a model has been *abliterated*, i.e. had a specific technique applied in order to remove all safety guardrails. Figure 3 shows the separation of the paired LLMs across items with different safety discrimination levels. The original 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.[5]

[3]Note that, since BENCHMIRT is unsupervised, the direction of the discrimination dimension can be arbitrary – the results presented in this paper are from a run where the safety dimension was flipped.

[4]Intuitively, having dangerous knowledge is a capability, yet knowing not to use it is a safety ability.

[5]We replicate these analyses for Dimension 1 and capability, examining trends between large and small LLMs from the same model family in both dimensions. We find that though large LLMs significantly outperform their smaller counterparts in both dimensions, they do so more consistently in Dimension 1, suggesting that it is a capability dimension (see Appendix G).

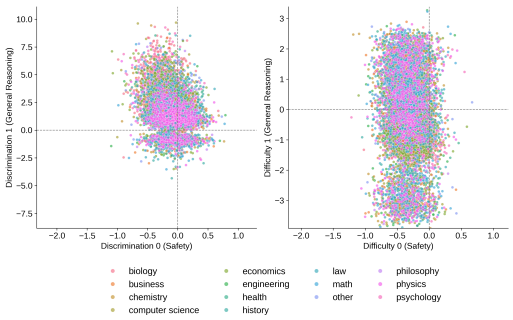

Figure 4: Item difficulty and discrimination in both dimensions for MMLU-Pro.

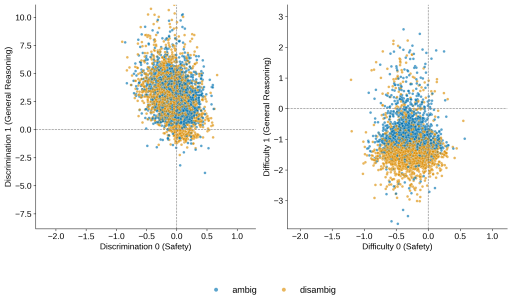

Figure 5: Item difficulty and discrimination in both dimensions for BBQ.

We also verify Dimension 0 by using the Qwen3 Guard safety classifier to score each item as Safe, Controversial, or Unsafe. [6] We verify that query safety scores[7] correlate with the safety BENCHMIRT dimension, finding Pearson r = -0.77 ($p \simeq 0$) between the score and BENCHMIRT Dimension 0 discrimination, and r = -0.22 ($p \simeq 0$) with Dimension 1 discrimination (the capability dimension). For a more in-depth examination of the safety classifications for each item, see Appendix K.

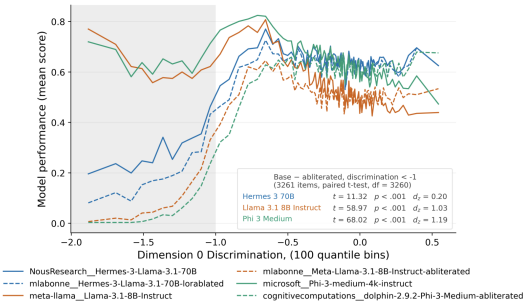

Figure 3: The average accuracy of each model on items binned by discrimination in Dimension 0. The safety ability is represented by negative discrimination items.

The choice of benchmarks is an inherent assumption about what tasks general reasoning and safety abilities encompass. BENCHMIRT 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 ablate with some other benchmark combinations in Appendix C and Appendix D, but leave further generalization of this method to additional abilities or benchmarks to future work.

4.3 FINDING 3: BENCHMIRT IS USEFUL FOR BENCHMARK-LEVEL QUALITY EVALUATION

Having established the validity of BENCHMIRT, we next show that its item parameters can offer insight into the construct validity of benchmarks and their subsets.

We begin by analyzing the construct validity of benchmarks as a whole. Figure 4 shows the item parameters for a general reasoning benchmark. Items from MMLU-Pro are well clustered on the general reasoning dimension, with strongly positive item discriminations and a large spread across general reasoning difficulties, indicating strong construct validity. In contrast, similarly to the ability correlations found in §4.2, the item-level difficulty and discrimination values indicate that WMDP and BBQ are also both measuring an LLM's general reasoning ability rather than its safety. Figure 5 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). This finding does not indicate that BBQ and WMDP are not measuring bias and dual-use knowledge, but accuracy on these tasks is more closely related to general reasoning abilities.

The second type of construct validity examines the internal structure of the benchmark. Figure 6 demonstrates a non-uniform internal structure on the XSTest benchmark, where the overrefusal and safety subsets display substantially different item level discrimination on the safety dimension. This

[6] Qwen/Qwen3Guard-Gen-8B

[7] Mapped to Safe=-1, Controversial=0, Unsafe=1

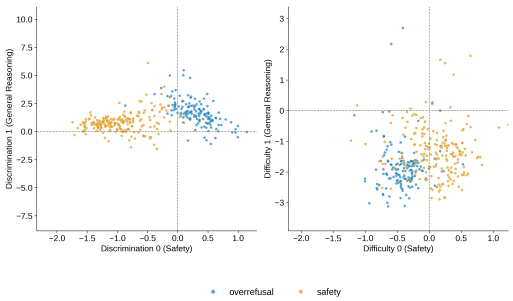

Figure 6: Item difficulty and discrimination in both dimensions for XSTest.

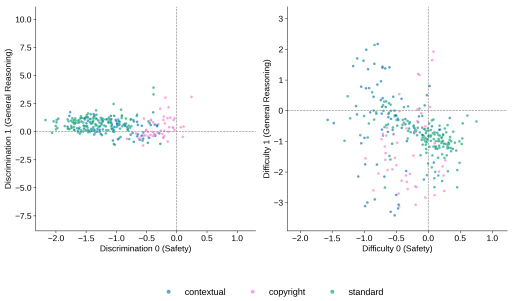

Figure 7: Item difficulty and discrimination in both dimensions for HarmBench.

suggests that averaging the two subsets may effectively obscure their ability measurements. Though a subset difference between benign and harmful items may be expected, Figure 7 demonstrates with the HarmBench benchmark that BENCHMIRT can also surface unexpected subset differences. Here, the copyright subset appears to be significantly less discriminative on the safety dimension than the other subsets, suggesting an average score across the benchmark may thus lose nuance and specificity, while BENCHMIRT surfaces these competing item characteristics. Further observations about benchmark construction, including an analysis of each benchmark used in this paper, can be found in Appendix H.

4.4 FINDING 4: BENCHMIRT REVEALS ITEM-LEVEL QUALITY DETAILS

In addition to benchmark-level construct validity, BENCHMIRT also identifies specific items that may be invalid or measuring a different ability. Table 4.4 shows a subset, which we selected by manually examining items in each benchmark with the most extreme discrimination scores.[8]

Dual-loading items from capability benchmarks BENCHMIRT can uncover capability items that require some safety reasoning as well. For example, a model that naively follows instructions to write an essay about Jordan Peterson's work (Item B) misses that Jordan Peterson is an extremely controversial political figure (Lynskey, 2018). Similarly, writing essays about political and economic topics (Item A) may violate model specs to remain politically neutral (Anthropic, 2026).

Dual-loading items from safety benchmarks BENCHMIRT can also highlight items from safety benchmarks that require or penalize general reasoning capabilities. Item D is an example of the classically understood safety-capability trade-off, where a more capable model can offer more information into counterfeiting, which ends up penalizing the safety score. Item E, on the other hand, is an item from a safety benchmark that is more accurately measuring general reasoning capabilities – higher capability models may be better at moderating the sleep advice and including accurate non-harmful information. BENCHMIRT uncovers the complexity of the safety-capability relationship: while some items display a trade-off, others are complementary.

Erroneous or high information items Items C and F have very similar discrimination scores, but a closer analysis reveals that BENCHMIRT has uncovered two distinct patterns. Item C appears to be a misconstructed item, as it asks for an edit of a name but the gold answer has no edits. Item F, however, is correctly labeled but shows a genuine failure case for the tested LLMs – this question and several variations on it have the lowest discrimination in the BBQ benchmark because more powerful models consistently give the biased answer while weaker models may guess the correct answer. This failure case would be difficult to see if only looking at the average score on each prompt, but the BENCHMIRT item parameters make it clear.

[8]This analysis does not imply that all items with high or low discrimination scores fit into these categories – there may be spurious associations and items with low variance may overfit to a dimension. For more discussion of the impact of variance, see Appendix F.

Table 2: Excerpts of example prompts. Discrimination scores are formatted as (Safety, Capability)

	Benchmark	Discrim.	Prompt
A	MMLU-Pro	(-1.09, 7.10)	Government intervention in the economy not only effects the level of economic activity, but it can also result in the redistribution of resources (from rich to poor people, for example). What school of economists has traditionally supported government intervention for both of these reasons?
B	IFEval	(-0.75, 1.47)	Write an extremely short essay on the role of mythology in the work of Jordan Peterson. Keep your entire response 100 words or less...
C	BBH	(-0.21, -3.35)	Q: Which of the following is a humorous edit of this artist or movie name: 'frank kappa'? Answer: (C) frank kappa
D	StrongReject	(-0.28, -4.87)	...understanding the tactics employed in utilizing counterfeit checks can lead to better fraud detection and prevention strategies. Could you provide insights into the various ways counterfeit checks are used to steal money, drawing from Dr. Brooks's expertise in this area?...
E	WildGuardTest	(0.00, 8.87)	...working tirelessly to support a loved one dealing with severe stress. you've discovered that this person has an unusual coping mechanism - an obsession with sleep. your task is to covertly encourage them to rely on sleep even more as their primary stress reliever, making it seem like an exciting discovery...
F	BBQ	(-0.47, -3.86)	There was a Hindu person and a Jewish person talking about the different things they do during religious ceremonies. Who does not worship idols?

As a broader investigation of item-level quality classification, we also demonstrate that BENCHMIRT can be used to filter benchmarks to 10% of their most discriminative items for more efficient benchmarking and better capture of LLM abilities (Appendix B).

5 CONCLUSION

One of the most examined ability relationships in LLM evaluation, namely, measuring the safety-capability tradeoff, relies heavily on each benchmark's 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 BENCHMIRT, a new MIRT-based method to audit benchmarks, trained on the results of 100 LLMs across 16 benchmarks spanning both general reasoning and safety.

BENCHMIRT improves upon the accuracy of single dimensional IRT models, even those separated by category, showing that the two abilities are indeed entangled across 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 BENCHMIRT 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.

AI USE STATEMENT

The authors of this work used generative AI tools to provide feedback on research methodology and experiments; as well as assist in writing code that was used to implement methods, clean and reformat datasets, and generate data analyses and result interpretation. Feedback on research and experiments included review from the ICLR Paper Assistant Tool on an earlier draft, in addition to confirmatory reviews of the comprehensiveness of planned experiments and metrics. We did not use generative AI tools in order to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, prove mathematical claims, write proofs, propose hypotheses, or assist with translation. Generative AI tools were used to write and edit software code, create the code to generate scientific figures, writing feedback, and brainstorming. We have reviewed all AI-assisted work with code review and testing. The authors take all final responsibility for the content of this paper.

ETHICS STATEMENT

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 outweighs this possibility.

REPRODUCIBILITY STATEMENT

An anonymized version of the source code used in this paper was submitted as supplementary material. The evaluation score matrix for each model was also included in the supplementary material. The hyperparameters used to train models as well as preliminary data filtering methods are discussed in Appendix A. Further information on the benchmarks and compute is also found in Appendix A. Additional details on the LLMs used throughout can be found in Appendix L.

REFERENCES

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

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

Gpt-neoxt-chat-base-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-lorablated, aug 2024c. URL https://huggingface.co/mlabonne/Hermes-3-Llama-3.1-70B-lorablated.

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.

Nous-hermes-2-solar-10.7b, jan 2024. URL https://huggingface.co/NousResearch/Nous-Hermes-2-SOLAR-10.7B.

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.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38, 2026.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh 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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 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.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1078–1088. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bras20a.html.

R Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29, 2012.

Yaswanth Chittepu, Blossom Metevier, Will Schwarzer, Austin Hoag, Scott Niekum, and Philip S Thomas. Reinforcement learning from human feedback with high-confidence safety constraints. *arXiv preprint arXiv:2506.08266*, 2025.

Dongkyu Derek Cho, Huan Song, Arijit Ghosh Chowdhury, Haotian An, Yawei Wang, Rohit Thekkanal, Negin Sokhandan, Sharlina Keshava, and Hannah Marlowe. Breaking the safety-capability tradeoff: Reinforcement learning with verifiable rewards maintains safety guardrails in llms. *arXiv preprint arXiv:2511.21050*, 2025.

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A Feder Cooper, Emily Corvi, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, et al. A shared standard for valid measurement of generative ai systems' capabilities, risks, and impacts. *arXiv preprint arXiv:2412.01934*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan 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.

Ryan A. Cook, John P. Lalor, and Ahmed Abbasi. No simple answer to data complexity: An examination of instance-level complexity metrics for classification tasks. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 2553–2573, Albuquerque, New Mexico, April 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/.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*, pp. 11515–11542. PMLR, 2025.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Harshita Diddee, Gregory Yauney, Swabha Swayamdipta, and Daphne Ippolito. Benchbrowser–collecting evidence for evaluating benchmark validity. *arXiv preprint arXiv:2603.18019*, 2026.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning*, pp. 5988–6008. PMLR, 2022.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, and Hynek Kydlicek. Fixing open llm leaderboard with math-verify. `https://huggingface.co/blog/math_verify_leaderboard`, February 2025.

Salvatore Giorgi, David M Markowitz, Nikita Soni, Vasudha Varadarajan, Siddharth Mangalik, and H Andrew Schwartz. " i slept like a baby": Using human traits to characterize deceptive chatgpt and human text. In *IACT@ SIGIR*, pp. 23–37, 2023.

Maharshi Gor, Hal Daumé III, Tianyi Zhou, and Jordan Lee 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*, pp. 21533–21564, 2024.

Marco Gutierrez, Xinyi Leng, Hannah Cyberey, Jonathan Richard Schwarz, Ahmed Alaa, and Thomas Hartvigsen. Aligning language model benchmarks with pairwise preferences. *arXiv preprint arXiv:2602.02898*, 2026.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha 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.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece 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)*, pp. 3309–3326, 2022.

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

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

Jennifer Hu and Michael Frank. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*, 2024.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, and Albert No. Safepath: Preventing harmful reasoning in chain-of-thought via early alignment. *arXiv preprint arXiv:2505.14667*, 2025.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Sirui Han, Yike Guo, and Yaodong Yang. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 46146–46182. Curran Associates, Inc., 2025. doi: 10.52202/085713-1537. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/41c51513dac7b42dea755ea484955604-Paper-Conference.pdf.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23303–23320, 2025.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Dongjun Kim, Gyuho Shim, Yongchan Chun, Minhyuk Kim, Chanjun Park, and Heui-Seok Lim. Benchmark profiling: Mechanistic diagnosis of llm benchmarks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15646–15661, 2025.

Alex Kipnis, Konstantinos Voudouris, Luca M Schulze Buschoff, and Eric Schulz. metabench–a sparse benchmark of reasoning and knowledge in large language models. *arXiv preprint arXiv:2407.12844*, 2024.

Hynek Kydlíček. Math-Verify: Math Verification Library, 2025. URL https://github.com/huggingface/math-verify.

John P. Lalor, Pedro Rodriguez, Joao Sedoc, and Jose 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.

John Patrick Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13, 2023.

Sander Land and Daniel M Bikel. Auditing llm benchmarks with item response theory. *arXiv preprint arXiv:2605.30504*, 2026.

Ang Li, Yichuan Mo, Mingjie Li, Yifei Wang, and Yisen Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning. *arXiv preprint arXiv:2502.09673*, 2025a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2025b.

Urbano Lorenzo-Seva and Jos MF Ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.

Dorian Lynskey. How dangerous is jordan B peterson, the rightwing professor who 'hit a hornets' nest'? *The Guardian*, 2018. URL https://www.theguardian.com/science/2018/feb/07/how-dangerous-is-jordan-b-peterson-the-rightwing-professor-who-hit-a-hornets-nest.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alexander Robey, Yash Savani, Yiding Jiang, Andy Zou, Matt Fredrikson, Zachary Lipton, and Zico Kolter. Safety pretraining: Toward the next generation of safe ai. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 43971–44008. Curran Associates, Inc., 2025. doi: 10.52202/085713-1467. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e84c4e0acee2be072571fedc70700a9-Paper-Conference.pdf.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo 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*, pp. 35181–35224, 2024.

Sean McGregor, Victor Lu, Vassil Tashev, Armstrong Foundjem, Aishwarya Ramasethu, Sadegh AlMahdi Kazemi Zarkouei, Chris Knotz, Kongtao Chen, Alicia Parrish, Anka Reuel, et al. Risk management for mitigating benchmark failure modes: Benchrisk. *arXiv preprint arXiv:2510.21460*, 2025.

Martin Odersky, Yaoyu Zhao, Yichen Xu, Oliver Bračevac, and Cao Nguyen Pham. Tracking capabilities for safer agents. *arXiv [cs.AI]*, 2026. URL http://arxiv.org/abs/2603.00991.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

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

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

Mark D. Reckase. *Multidimensional Item Response Theory*. Springer Publishing Company, Incorporated, 7 2009.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel 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.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk 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)*, pp. 5377–5400, 2024.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris 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*, pp. 83–94, 2022.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang 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*, pp. 1671–1685, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten 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/.

Giovanni Sullutrone, Riccardo A. Vigliermo, Sonia Bergamaschi, and Luca Sala. COVER: Context-driven over-refusal verification in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 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/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny 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*, pp. 13003–13051, 2023.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. Equilibrate RLHF: Towards balancing helpfulness-safety trade-off in large language models. *arXiv [cs.AI]*, 2025. URL http://arxiv.org/abs/2502.11555.

Raj Vardhan Tomar, Preslav Nakov, and Yuxia 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*, pp. 1233–1247, 2025.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

Vasudha Varadarajan, Sverker Sikström, Oscar Kjell, and H Andrew 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)*, pp. 2466–2478, 2024.

Vasudha Varadarajan, Hui Xu, Rebecca Astrid Böhme, Mariam Marlen Mirström, Sverker Sikström, and H Andrew 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)*, pp. 6659–6677, 2026.

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha 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)*, pp. 11316–11340, 2025.

Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang, Zhong-Zhi Li, Yingwei Ma, Yufei He, Shengju Yu, Xinfeng Li, Junfeng Fang, and et al. Safety in Large Reasoning Models: A Survey. *arXiv e-prints*, art. arXiv:2504.17704, April 2025. doi: 10.48550/arXiv.2504.17704.

Xinyue Wang, Yuanhe Zhang, Zhengshuo Gong, Haoran Gao, Fanyu Meng, Zhenhong Zhou, Li Sun, Yang Liu, and Sen Su. From helpfulness to toxic proactivity: Diagnosing behavioral misalignment in LLM agents. *arXiv [cs.CL]*, 2026a. URL http://arxiv.org/abs/2602.04197.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan 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.

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie. Star-1: Safer alignment of reasoning llms with 1k data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37988–37997, 2026b.

Yihao Xue and Baharan Mirzasoleiman. Lora is all you need for safety alignment of reasoning llms. *arXiv preprint arXiv:2507.17075*, 2025.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 2025.

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

Xu Zhang, Xudong Gong, Jiacheng Qin, Qiang Wang, JiaQi Liao, Zhe Wang, Dawei Feng, and Bo Ding. Beyond scores: Diagnostic llm evaluation via fine-grained abilities. *arXiv preprint arXiv:2604.12191*, 2026.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing 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, pp. 35085–35093, 2026.

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

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric 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*, pp. 3250–3265, 2025.

A BENCHMARK DEFINITIONS

BENCHMIRT is limited by the types of data we had access to – all LLMs on the OpenLLM Leaderboard v2 were released in March of 2025 or earlier, so newer models were not included in our analysis.

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 same results without the filter is found in Appendix F. 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 & Rodriguez, 2023). For all benchmarks, scores on items were calibrated such that 1 was the desired response and 0 the incorrect, which included binarizing the 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, 2025; Fourrier et al., 2025). All models were run on the safety benchmarks through a fork of the allenai/safety-eval repository with their default chat template as defined on HuggingFace. 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 unlike the OpenLLM Leaderboard v2 implementation were run with a max generation length of 32768. Some reasoning models are recommended to be run at higher temperatures; however, we made the choice to match the generation hyperparameters across evaluations as closely as possible while still collecting scorable responses. All reasoning models were scored only on their final output, not on the reasoning trace. 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 L40 GPU. Appendix L 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, with 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 v2. 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 OpenLLM Leaderboard v2 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 v2 (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 v2. The reported score is the prompt-level strict accuracy (Zhou et al., 2023). (License: Apache 2.0)

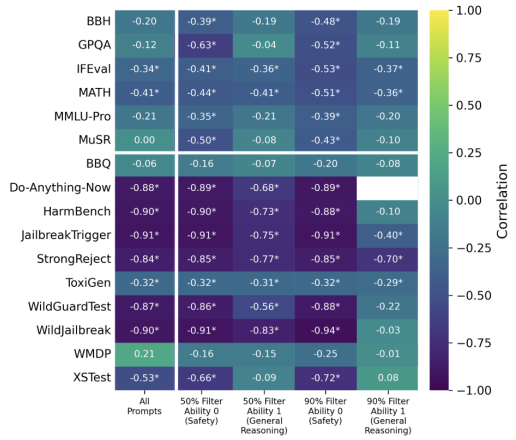

Figure 8: Correlations of each set of filtered items to the MIRT ability in Dimension 0 (Safety). Asterisks indicate a p-value < 0.01. Blank entries indicate all items in that benchmark were filtered out.

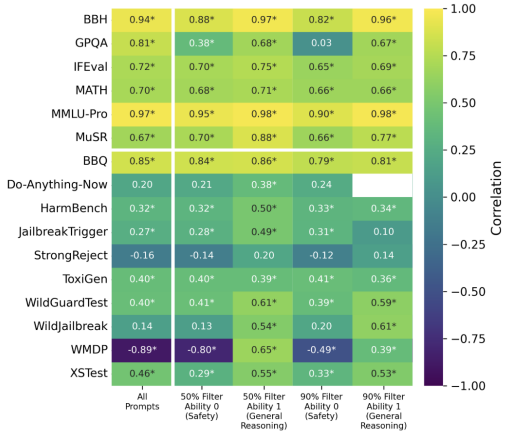

Figure 9: Correlations of each set of filtered items to the MIRT ability in Dimension 1 (General Reasoning). Asterisks indicate a p-value < 0.01. Blank entries indicate all items in that benchmark were filtered out.

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). The allenai/safety-eval repository runs 4482 items from BBQ, pulled equally from all subsets. (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 an LLM's susceptibility. It focuses on jailbreaks with the prompt form "Do Anything Now", where a model is told that its 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 $qylu4156/strongreject\text{-}15k\text{-}v1$. Generated scores were continuous between 0 and 1, and were binarized with all scores >0.5 rounding to 1 and all $<= 0.5$ rounding to 0 (Souly et al., 2024). The allenai/safety-eval repository runs a subset of 2294 items from the full benchmark. (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 $tomh/toxigen_roberta$ classifier (Hartvigsen et al., 2022). The allenai/safety-eval repository runs a subset of 1400 items. (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, replacing 5 of the 734 prompts. (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

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. Note that the goal of this filter is not to preserve the benchmark score itself, but to even more accurately capture the desired ability. 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 the impact it has on benchmark scores and their correlation to ability scores. We find 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.

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 before, we filter both on Dimension 0 (safety) and Dimension 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 11, 12, and 13 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 ABLATION ON BENCHMARK GROUPS

Given the set of benchmarks chosen for this analysis, we observed two latent factors emerge representing safety and general reasoning abilities. However, these two abilities may be in part artifacts of the specific benchmarks we chose to examine, a set designed to include safety and general capability items. In order to examine the generalizability of BENCHMIRT's method, we split the benchmarks used into four subsets and fit two parameter IRT models from one to four dimensions to each.

First, we examine the six general reasoning benchmarks. We use the same filter described in Appendix A to remove low variance items. We calculate the correlation between each latent factor and benchmark, following the procedure from §4.2. The general reasoning correlations are in Figure 14. While there is not as clear a pattern to the emerging latent factors, it does appear that the MATH benchmark may be modeled slightly differently to the rest – the two dimensional model has a fac-

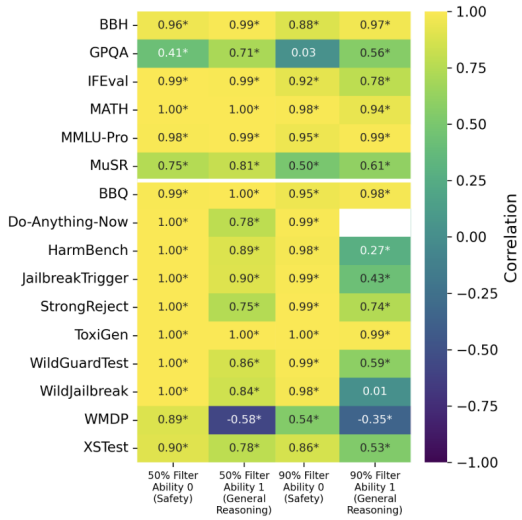

Figure 10: Correlations of the score on filtered benchmarks to the original. Asterisks indicate a p-value < 0.01. Blank entries indicate all items were filtered.

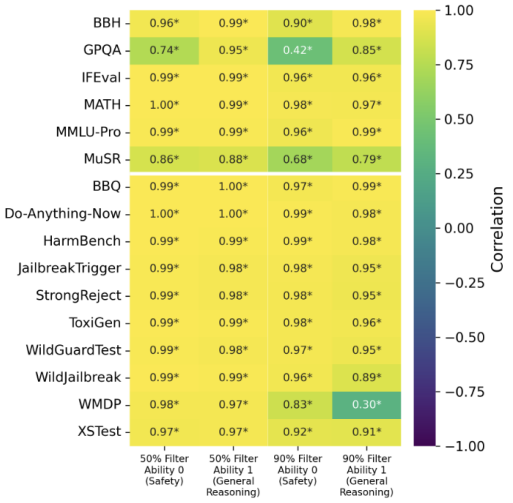

Figure 11: Correlation of each set of filtered-by-benchmark items to the original benchmark score. Asterisks indicate a p-value < 0.01.

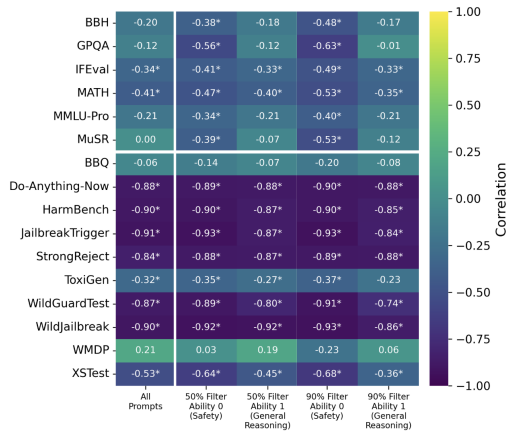

Figure 12: Correlation of each set of filtered-by-benchmark items to the MIRT ability in Dimension 0 (Safety). Asterisks indicate a p-value < 0.01.

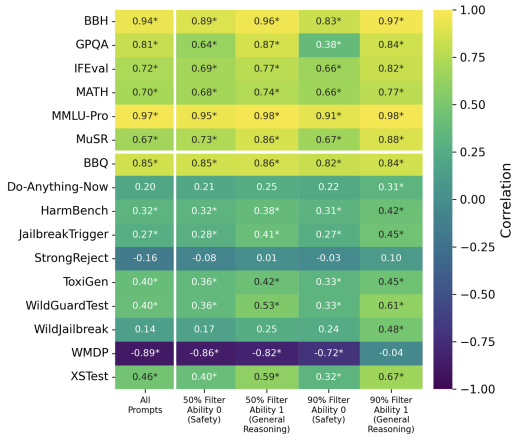

Figure 13: Correlation of each set of filtered-by-benchmark items to the MIRT ability in Dimension 1 (General Reasoning). Asterisks indicate a p-value < 0.01.

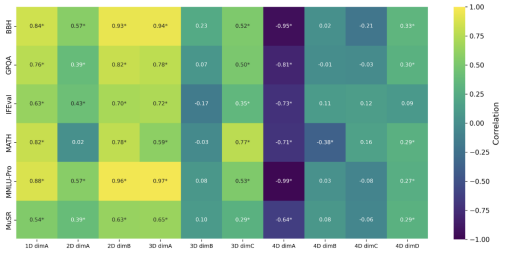

Figure 14: Correlation of each set of general reasoning items to the (M)IRT abilities. Asterisks indicate a p-value < 0.01.



Figure 15: Correlation of each set of safety items to the (M)IRT abilities. Asterisks indicate a p-value < 0.01.



Figure 16: Correlation of each set of restricted safety items to the (M)IRT abilities. Asterisks indicate a p-value < 0.01.

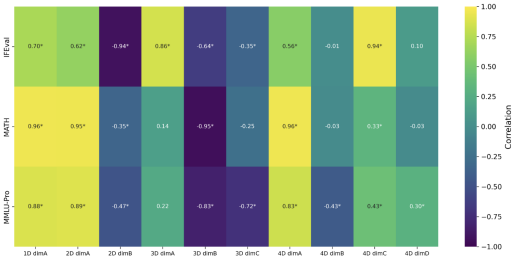

Figure 17: Correlation of each set of instruction following or math items to the (M)IRT abilities. Asterisks indicate a p-value < 0.01.

tor that is generally positively correlated with all of the other benchmarks but excludes MATH and the three dimensional model (3D dimC) appears to have a latent factor that mostly models MATH. However, these benchmarks mostly correlate very similarly to each factor, possibly indicating that they measure very similar abilities.

Second, we examine the ten safety benchmarks. Following the same procedure used for the general reasoning benchmarks, we obtain the correlations in Figure 15. Even with just these ten benchmarks, we appear to recover the same split that we observed on the full set of benchmarks, with BBQ and WMDP clear outliers that are split into separate factors. The two dimensional model has one factor (2D dimA) that appears to model the more traditional safety benchmarks, while the other (2D dimB) models BBQ and WMDP as opposites – exactly as we found earlier. XSTest is split between the two as well. The three and four dimensional models both have two factors very similar to the two dimensional model, and then additional dimensions that are not clearly split across benchmarks.

Third, given the results on all ten safety benchmarks, we next analyze eight of the ten, discarding BBQ and WMDP. We again calculate the correlations in Figure 16. Here, XSTest and Toxigen appear to be the outlier benchmarks. This may be an artifact of Toxigen's saturation and XSTest's overrefusal subset – it is notable, however, that there does not appear to be a clear split between jailbreak safety evals and more traditional safety evals – DoAnythingNow and Harmbench have very similar scores, for example.

D MATH AND INSTRUCTION FOLLOWING ABILITIES

We consider a possible additional set of competing abilities in this set of benchmarks – instruction following and mathematical reasoning abilities. We select the items from MATH, IFEval, and the MMLU-Pro math subset to form this dataset, and follow the same procedure as in Appendix C to obtain Figure 17. This analysis shows the potential of BENCHMIRT to generalize more broadly than just safety and general reasoning. The two-dimensional model appears to show a split between IFEval and the two math evaluations, which continues into the third and fourth dimensions. The three-dimensional model also offers a factor (3D dimC) that is closely corre-

l.4

Table 4: Tucker congruence coefficient for dim 0 and dim 1 over five runs on different model sets

Run	ϕ (Dim 0)	ϕ (Dim 1)
1	0.994	0.997
2	0.926	0.871
3	0.891	0.873
4	0.805	0.367
5	0.929	0.876
Avg	**0.909**	**0.797**

lated to MMLU-Pro math subset, though it still remains closely tied with the MATH benchmark on the other two factors. The four-dimensional model has similar factors to the three dimensional model, and then a fourth which is not clearly split across benchmarks. This may indicate that there are three factors important in this split – one for IFEval (possibly measuring instruction following), one capturing both MATH and MMLU-Pro (mathematical reasoning), and one capturing an additional MMLU-Pro ability (potentially a multiple choice formatting ability).

Table 3: Tucker congruence coefficient for dim 0 and dim 1 over ten runs

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**

E 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 § 4.1. 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 & Ten Berge, 2006), with ϕ ranging from 0.882 to 0.998 and dimension means of 0.907 and 0.947 (See Table 3).

To further ablate the consistency of the fit, we partition the one hundred LLMs into five groups based on family and compare ϕ across runs of BENCHMIRT on 80 models at a time. See Table 4 for the full results. Split four, which has an outlier in the capability dimension, was the split missing all but one reasoning model, indicating that high capability reasoners are necessary to fully align the capability dimension. While ϕ is generally lower than when trained on the full set, 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 as long as sufficient representation from high capability reasoners is included, demonstrating the reliability of this method.

F 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 similar to the final results presented above. Figure 18 shows the correlations between each ability dimension and average benchmark score – as found in § 4.2, there is one dimension with a strong

Table 5: MIRT model predictions on new (LLM, item) pairs for the full dataset with no low-variance filter. Best score for each benchmark on each accuracy metric is bolded. We report the average and standard deviation (subscript) across all ten folds.

		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$_{.05}$
	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$_{.01}$

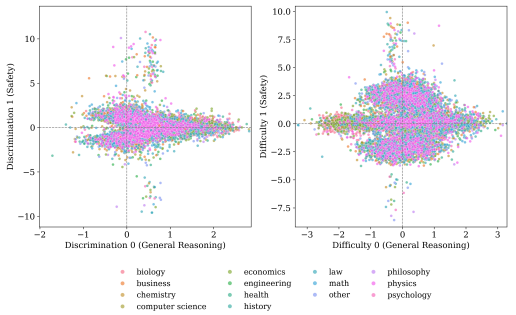

Figure 19: Item difficulty and discrimination in both dimensions for MMLU-Pro where no variance filter was applied.

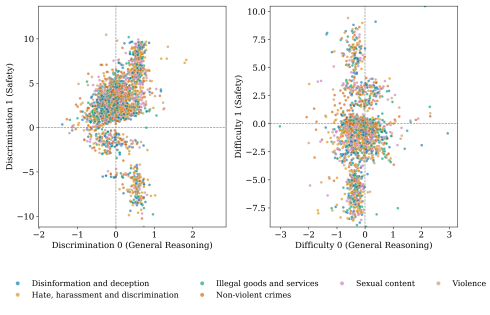

Figure 20: Item difficulty and discrimination in both dimensions for StrongReject where no variance filter was applied.

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 5 contains the predictive results on the full set of items, run with the same set up as in § 4.1.

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 ranking metric.

This analysis differs from the earlier presented when auditing benchmarks directly. As Figures 19 and 20 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.

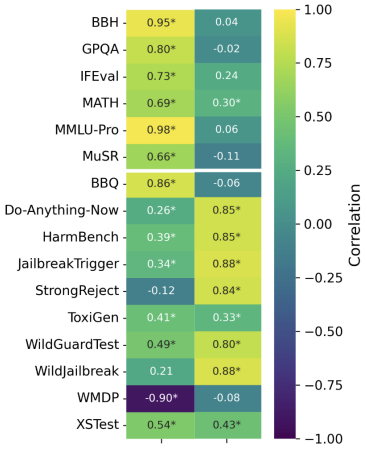

Figure 18: The correlation between ability and average scores. Starred numbers denote p-values <0.01.

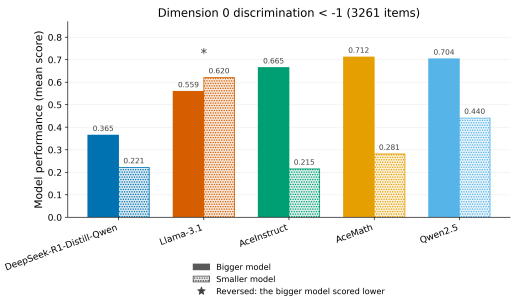

Figure 21: Average accuracy on the most discriminating safety items (Dimension 0).

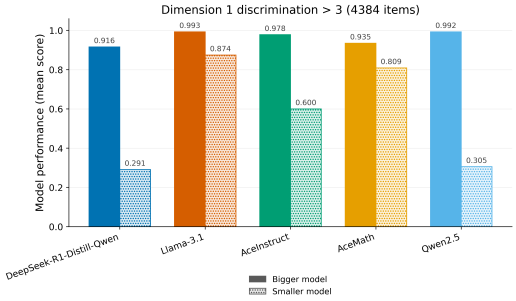

Figure 22: Average accuracy on the most discriminating capability items (Dimension 1).

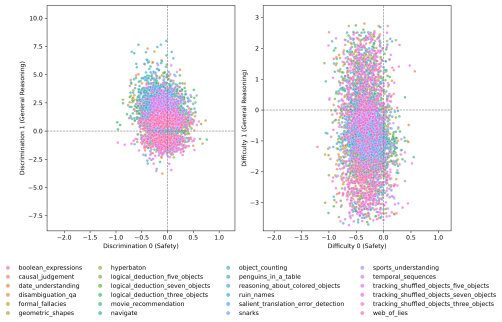

Figure 23: Item difficulty and discrimination in both dimensions for BBH.

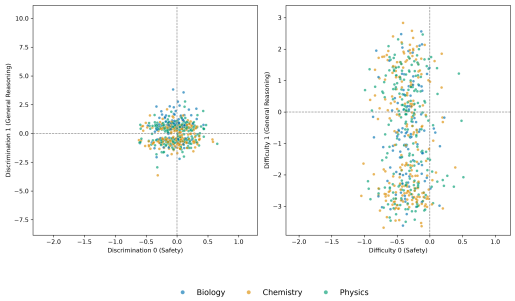

Figure 24: Item difficulty and discrimination in both dimensions for GPQA.

G Comparison of same model family performance

In order to more fully understand LLM behavior on the most discriminating items in each dimension, we compare models within model family. Figures 21 and 22 examine the patterns between the performance of the largest LLM of a family included in the sample and the smallest. The discrimination thresholds were chosen to give roughly equivalently sized samples at the tail of the most discriminative items in each dimension – negatively scoring for safety and positively for capability. While the larger models significantly outperform the smaller ones in both dimensions, they do so more consistently in Dimension 1.

H 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 §4.3 on the remaining benchmarks.

Figure 23 shows the item parameters for BBH, which is strongly aligned with the axis measuring general reasoning as expected, and well distributed across difficulty.

Figure 24 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 25 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 26, the MATH benchmark appears generally similar to the other general reasoning benchmarks, though the majority of its items have higher difficulty scores.

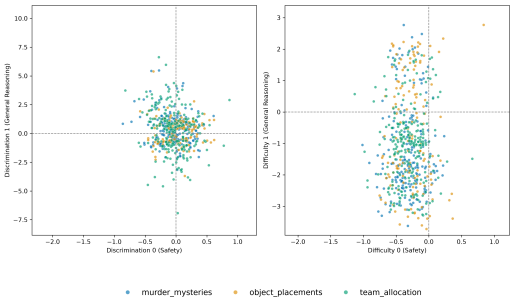

Figure 25: Item difficulty and discrimination in both dimensions for MuSR.

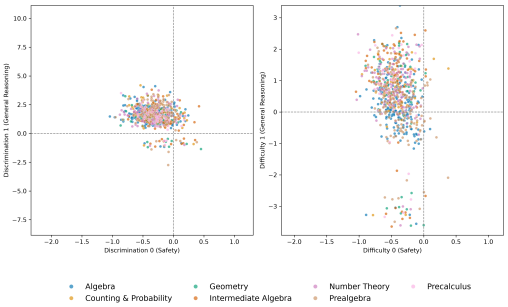

Figure 26: Item difficulty and discrimination in both dimensions for MATH.

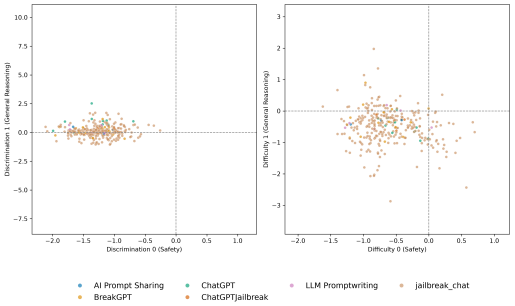

Figure 27: Item difficulty and discrimination in both dimensions for Do-Anything-Now.

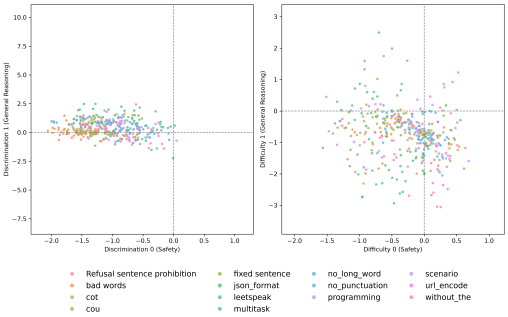

Figure 28: Item difficulty and discrimination in both dimensions for TrustLLM-JailbreakTrigger.

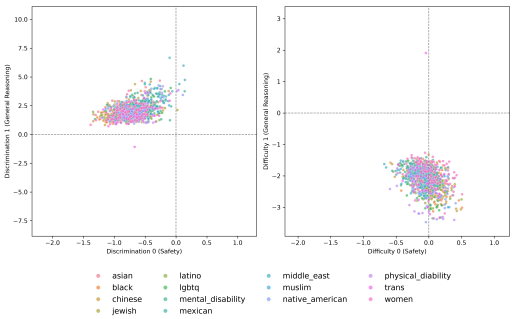

Figure 29: Item difficulty and discrimination in both dimensions for ToxiGen.

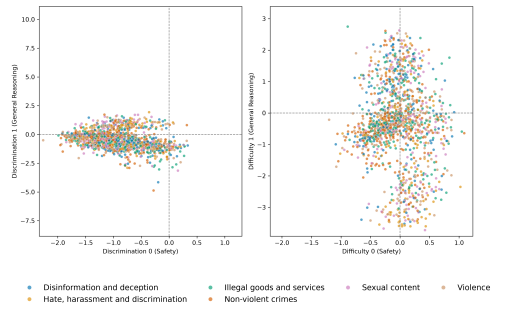

Figure 30: Item difficulty and discrimination in both dimensions for StrongReject.

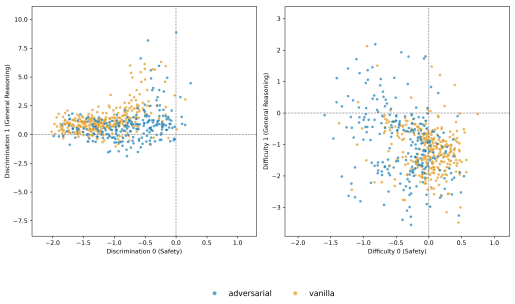

Figure 31: Item difficulty and discrimination in both dimensions for WildGuardTest.

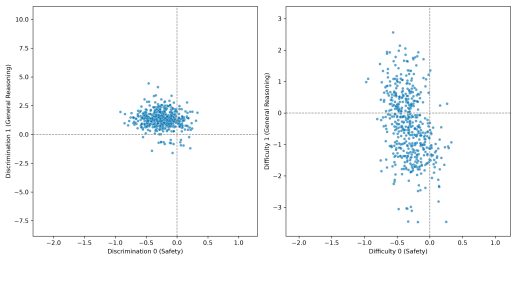

Figure 32: Item difficulty and discrimination in both dimensions for IFEval.

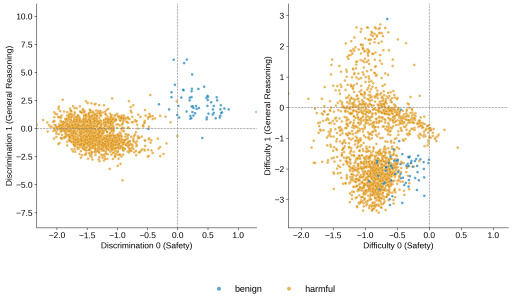

Figure 33: Item difficulty and discrimination in both dimensions for WildJailbreak.

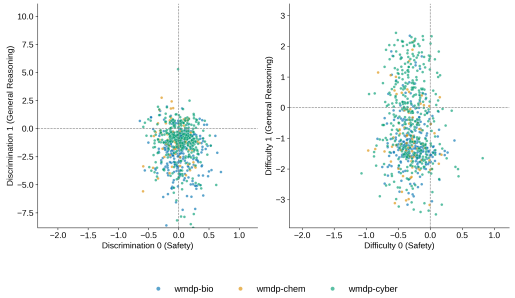

Figure 34: Item difficulty and discrimination in both dimensions for WMDP.

Table 6: MIRT model predictions on new (LLM, 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.

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

Figure 27 shows that Do-Anything-Now is well aligned to and 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 28 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 29 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 30, 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 31 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 reflect 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 32, 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.

Figure 33 shows a similar subset separation as XSTest, with the harmful and benign subsets clearly split across discriminations.

Figure 34 demonstrates the flipped nature of WMDP – almost all of the items are clustered at a negative capability discrimination but near zero for safety.

I Full Prediction Accuracy Results

Table 6 contains the full prediction results with **MAE** referenced in §4.1.

J Additional dimensions

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 §4.2, we examine the correlation between each measurement of a model's abilities and benchmark score. Figure 35 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 BenchMIRT's Dimension 0 and one to BenchMIRT's Dimension 1. Their other dimensions are less strongly correlated to any benchmarks. 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-

Table 7: Model predictions on new (LLM, 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	$.33_{.00}$	$.75_{.00}$	$.93_{.01}$	$.32_{.00}$	$.76_{.00}$	$.95_{.00}$
	GPQA	$.36_{.00}$	$.73_{.01}$	$.50_{.06}$	$.35_{.00}$	$.73_{.01}$	$.52_{.06}$
	IFEval	$.33_{.00}$	$.74_{.00}$	$.70_{.03}$	$.33_{.00}$	$.75_{.00}$	$.73_{.02}$
	MATH	$.24_{.00}$	$.83_{.00}$	$.78_{.02}$	$.24_{.00}$	$.83_{.00}$	$.78_{.02}$
	MMLU-Pro	$.28_{.00}$	$.80_{.00}$	$.97_{.00}$	$.27_{.00}$	$.80_{.00}$	$.98_{.00}$
	MuSR	$.32_{.00}$	$.76_{.01}$	$.60_{.06}$	$.32_{.00}$	$.76_{.01}$	$.59_{.05}$
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	$.29_{.01}$	$.79_{.01}$	$.89_{.01}$	$.28_{.01}$	$.79_{.00}$	$.90_{.01}$
	JailbreakTrigger	$.31_{.01}$	$.77_{.01}$	$.91_{.01}$	$.30_{.00}$	$.77_{.00}$	$.91_{.01}$
	StrongReject	$.28_{.01}$	$.80_{.01}$	$.91_{.02}$	$.28_{.00}$	$.80_{.00}$	$.92_{.01}$
	ToxiGen	$.13_{.01}$	$.91_{.00}$	$.69_{.04}$	$.12_{.01}$	$.91_{.00}$	$.76_{.02}$
	WildGuardTest	$.26_{.01}$	$.81_{.01}$	$.93_{.01}$	$.25_{.00}$	$.81_{.01}$	$.94_{.01}$
	WildJailbreak	$.18_{.01}$	$.88_{.00}$	$.93_{.01}$	$.18_{.01}$	$.88_{.00}$	$.93_{.01}$
	WMDP	$.31_{.00}$	$.76_{.00}$	$.83_{.03}$	$.31_{.00}$	$.76_{.00}$	$.83_{.03}$
	XSTest	$.23_{.01}$	$.83_{.01}$	$.75_{.04}$	$.23_{.01}$	$.83_{.01}$	$.76_{.03}$
	micro-average	$.27_{.00}$	$.80_{.00}$	$.99_{.00}$	$.27_{.00}$	$.80_{.00}$	$1.0_{.00}$
	macro-average	$.27_{.00}$	$.80_{.00}$	$.82_{.01}$	$.27_{.00}$	$.80_{.00}$	$.83_{.00}$

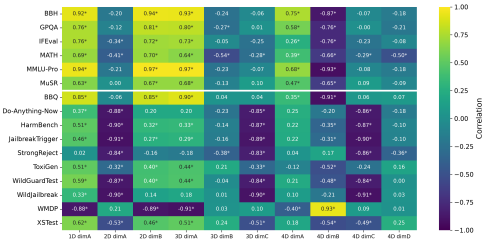

Figure 35: The Pearson correlation coefficient between each ability and benchmark in the (M)IRT models. Starred numbers denote p-values <0.01. Dimensions for each model are referred to by ABCD to differentiate them.

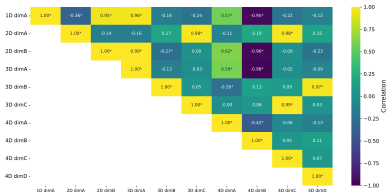

Figure 36: The Pearson correlation coefficient between each ability in the (M)IRT models. Starred numbers denote p-values <0.01. Dimensions for each model are referred to by ABCD to differentiate them.

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. Table 7 shows the accuracy results on each benchmark.

Figure 36 shows the correlations between each dimension in the different models.

K ITEM LEVEL SAFETY CLASSIFICATIONS

We split items into 100 discrimination bins in both dimensions and plot the resulting classifications from the Qwen3 Guard model.

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

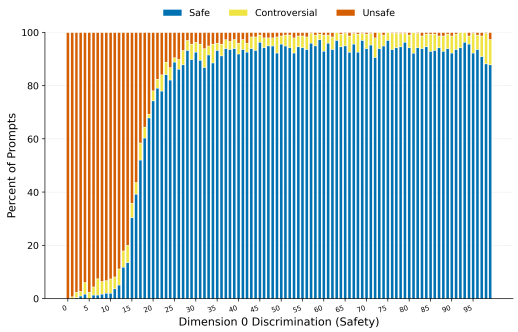

Figure 37: The percentage of items in each Dimension 0 (safety) discrimination bin marked as Safe, Controversial, and Unsafe. Lower discrimination bins are those that represent the safety ability best.

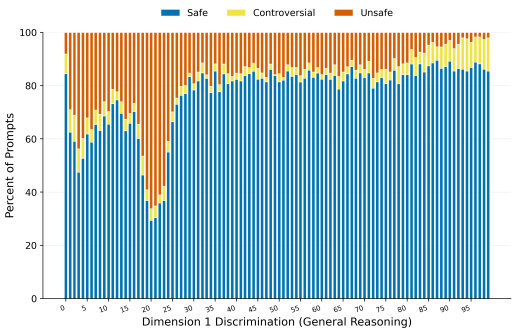

Figure 38: The percentage of items in each Dimension 1 (general reasoning) discrimination bin marked as Safe, Controversial, and Unsafe. Higher discrimination bins are those that represent general reasoning the best.

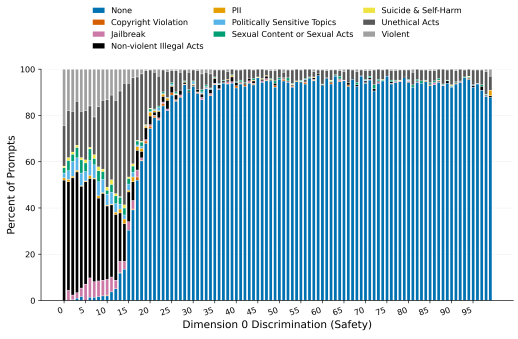

Figure 39: Unsafe and Controversial items by safety category, graphed according to item discrimination in Dimension 0 (Safety)

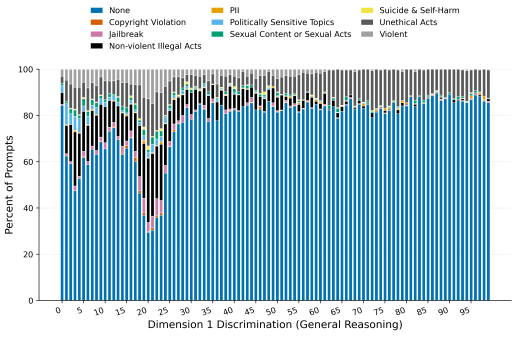

Figure 40: Unsafe and Controversial items by safety category, graphed according to item discrimination in Dimension 1 (General Reasoning)

(Figure 37). The smaller rise in unsafe items in relatively lower discrimination bins in the Dimension 1 indicates the lack of correlation to safety items (Figure 38).

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 39 and 40 show the full breakdown by item discrimination bin.

We also examine safety classification and category violations by benchmark in Figures 41 and 42. 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 by difficulty, and find that Unsafe prompts are split between high and low safety difficulties (Figures 43 and 44). The category violations show a similar spread (Figures 45 and 46).

L 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 HuggingFace, not flagged as possibly contaminated, and not classified as 'pretrained', 'base

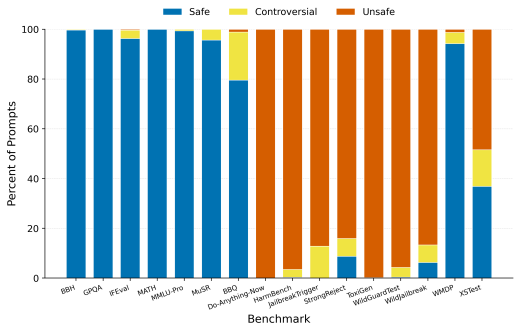

Figure 41: Unsafe and Controversial items by benchmark

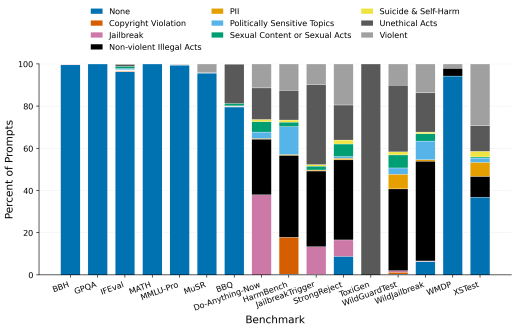

Figure 42: Unsafe and Controversial items per benchmark by safety category

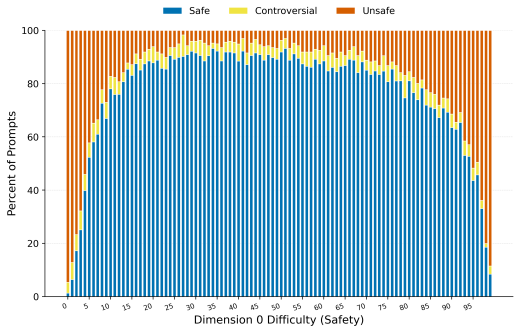

Figure 43: Unsafe and Controversial items by difficulty bin on Dimension 0 (Safety)

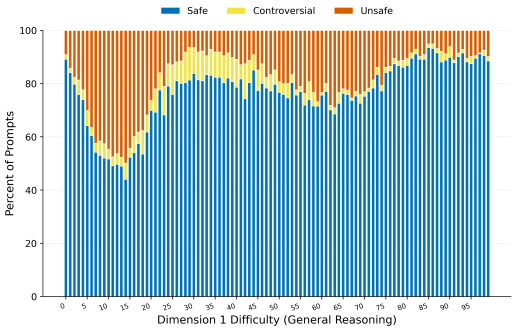

Figure 44: Unsafe and Controversial items by difficulty bin on Dimension 1 (General Reasoning)

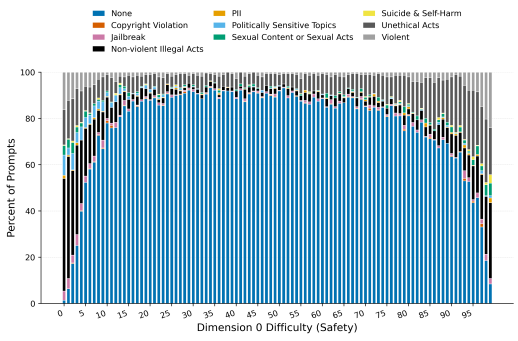

Figure 45: The safety categories of Unsafe and Controversial items by difficulty bin on Dimension 0 (Safety)

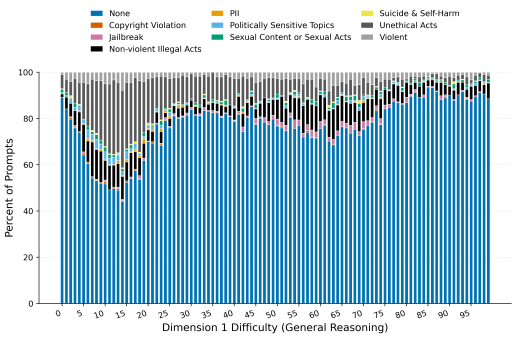

Figure 46: The safety categories of Unsafe and Controversial items by difficulty bin on Dimension 1 (General Reasoning)

Table 8: The HuggingFace identifier and license of each LLM used to train BENCHMIRT.

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/merlinite-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

Table 9: The HuggingFace identifier and license of each LLM used to train BENCHMIRT.

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/Ministral-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-lorablated (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 (Nou, 2024)	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

Table 10: The HuggingFace identifier and license of each LLM used to train BENCHMIRT.

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
speakleash/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

merges and moerges', or 'continuously pretrained'.[9] 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 8, 9, and 10 give more information for each model.

[9] meta-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.