Rethinking Math Benchmarks for LLMs using IRT

Jane Castleman* JANEEC@PRINCETON.EDU
Nimra Nadeem* NNADEEM@PRINCETON.EDU
Tanvi Namjoshi* NAMJOSHI@PRINCETON.EDU
Lydia T. Liu LTLIU@PRINCETON.EDU
35 Olden Street, Princeton, NJ 08544

Abstract

Several datasets have been created to evaluate LLM performance on mathematical reasoning tasks. Performance on these benchmarks is used as a proxy for a model's math ability and to rank their capability relative to other models. These rankings play a crucial role for AIEd practitioners in selecting models for applications like math tutoring. Recent research has argued that several of these benchmarks have become too saturated, prompting the creation of new datasets with more difficult tasks. How can we gauge the effectiveness of these benchmarks for measuring math skills and producing reliable rankings? Leveraging the psychometric framework of Item Response Theory, we examine three math benchmarks: GSM8K, MATH, and MathOdyssey. We find that GSM8K and MathOdyssey are not suited to properly evaluate the current range of frontier model abilities, and are instead suited to models with lower and higher math abilities respectively. Moreover, current rankings derived from these benchmarks are unstable and fail to reliably capture the latent math ability they aim to measure. To remedy these issues, we recommend the integration of IRT analysis into the process of selecting questions for future benchmarks.

Keywords: Item Response Theory, LLM Benchmarks, AIEd, LLM Evaluation

1. Introduction

Benchmarks have become ubiquitous tools for language model developers to communicate the level of a model's ability on a set of tasks. They are often used to compare frontier models and argue that a new model is "state-of-the-art," evidenced by higher performance. Burnell et al. (2023) argue that benchmarks fail for two key reasons: they quantify performance in aggregate metrics and do not provide insight into precise failure instances, making it difficult to understand the specific capabilities models actually possess. Furthermore, as model abilities increase, benchmarks become outdated (Yu et al., 2023), and there are growing concerns of training set contamination (Li, 2023). For example, the popular GSM8K benchmark used to test model ability on grade school math, sees extremely high levels of performance,[1] with top models earning over 90% accuracy. On the other hand, the MathOdyssey (Fang et al., 2024) dataset, developed to address performance saturation on GSM8K and training set contamination, sees a top accuracy of only 65%.[2]

Given the limitations of current benchmarks, we are interested in answering the following questions: **(1)** To what extent do current benchmarks robustly estimate and rank the

* These authors contributed equally to this work
1. See https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k
2. See https://mathodyssey.github.io/

abilities of LLMs? **(2)** How can we build benchmarks that better inform us of model abilities, even when model abilities increase?

To answer these questions, we use Item Response Theory (IRT) (Hambleton and Swaminathan, 1985), a common educational and psychometric statistical framework, to evaluate benchmarks for math reasoning in LLMs. Because AIEd systems depend on LLMs' math ability to explain math concepts to students (Xu et al., 2024; Wang et al., 2024), it is crucial to reliably estimate model math ability.

However, several AIEd companies are forced to develop their own benchmarks or to rely on a small set of informal questions to evaluate their products since prominent benchmarks are not designed for educational tasks (Wang et al., 2024; Miller and DiCerbo, 2024). As more benchmarks for AIEd applications are created, our framework provides a method to test if they effectively estimate the ability of models. Our work offers three main contributions:

- Extract model response patterns for math tasks to create a large test-taking population of LLMs in order to leverage IRT
- Quantitatively evaluate the validity and effectiveness of popular math benchmarks to discriminate between the abilities of frontier models
- Propose a framework for selecting individual questions that are most effective in discriminating between model ability, which can be generalized to future AIEd benchmarks

Overall, this framework will be crucial for the valid and effective evaluation of AIEd system abilities, since it not only evaluates the validity of current benchmarks but also facilitates the design and evaluation of future benchmarks.

2. Related Work

Our work is related to two key areas of research: creation and evaluation of LLM benchmarks, particularly in the development of AIEd tools, and use of IRT in NLP research.

2.1. Evaluating LLM Benchmarks

LLM benchmarks aim to standardize the evaluation of rapidly advancing state-of-the-art models. However, despite the popularity of benchmarks, researchers are calling into question their efficacy for understanding LLM capabilities (Anwar et al., 2024; Burnell et al., 2023). Ramesh et al. (2024) point out that because benchmarks evaluate LLMs on discrete tasks, they will inevitably fail to evaluate all tasks a model is capable of performing (Yu et al., 2023). Although benchmarks are appealing as a straightforward method to compare models, they have fundamental weaknesses and have not yet been rigorously validated.

Therefore, it is increasingly important to reliably assess benchmarks and identify which questions in a given benchmark are capable of providing a good understanding of model capabilities. Currently, there is a large demand for benchmarks to be created in the AI for Education space. Miller and DiCerbo (2024) introduces a new benchmark, COMTA, specifically for math education. However, given the challenges of creating tests for education-related abilities, developers of AIEd products often rely on ad-hoc examples or models that

perform well on standard math reasoning tasks. Thus, there is an opportunity to create benchmarks for education more mindfully. We turn to IRT to evaluate the validity of math benchmarks, but the analysis can extend to the creation of benchmarks in many high-impact domains.

2.2. IRT in NLP

IRT has become a foundational approach in educational testing due to its ability to model the interaction between individuals' latent traits and individual test item characteristics (Linden and Hambleton, 1997; Embretson and Reise, 2000; Chen et al., 2021). Lalor et al. (2016) apply IRT to create more challenging benchmarks for NLI tasks and Lalor et al. (2019), use response patterns from an artificial crowd of LSTM models to fit IRT parameters. Rodriguez et al. (2021) incorporate the difficulty and discrimination parameters of individual benchmark questions when ranking models on the SQuaD leaderboard, showing that this provides a better ranking than accuracy alone. Vania et al. (2021) compare IRT fits across several English NLU datasets, using response patterns from a range of BERT-based Transformer models. (Polo et al., 2024) employed IRT parameters to estimate LLM performance efficiently, demonstrating that approximately 100 curated samples could reliably approximate benchmark accuracy.

Our research expands on prior work in two key ways: **(1)** We use recent frontier models rather than traditional NLP or BERT-based models **(2)** We focus on math benchmarks due to their significance for the AIEd domain. To handle large modern data sets, we use a variational Bayesian inference algorithm for fitting IRT models, which has become standard practice in applications of IRT to NLP (Wu et al., 2020; Vania et al., 2021; Rodriguez et al., 2021; Lalor et al., 2019).

3. Methods

To evaluate the validity of math reasoning benchmarks, we fit 2-PL IRT models on response patterns from a population of models on a subset of the GSM8K, MATH and MathOdyssey benchmarks. We find our IRT models have strong model fit, indicating the validity of our IRT models in evaluating the discrimination and difficulty of these benchmarks.

3.1. Item Response Theory

Item Response Theory (IRT), a statistical framework widely used in educational and psychological testing due to its ability to model the interaction between an individual's latent ability and the characteristics of test items (Hambleton and Swaminathan, 1985; Chen et al., 2021). Unlike Classical Test Theory (CTT), which evaluates test performance at an aggregate level (Novick, 1966), IRT estimates parameters for each test item and individual respondent, enabling a more nuanced analysis.

We use the two-parameter logistic (2PL) model (Edwards, 2009), mathematically defined as follows:

$$P(x_{i,j} = 1|\theta_i, b_j, a_j) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{1}$$

For a benchmark B and set of models M, θ_i represents the latent ability of model $i \in M$, $x_{i,j}$ denotes model i's response to item $j \in B$, and b_j and a_j correspond to the item's difficulty and discrimination parameters, respectively. Estimating an IRT model involves deriving the optimal values for θ_i for each model i, and b_j and a_j for each test item j based on observed response patterns - i.e. whether a model answered a given item correctly or incorrectly. The 2PL IRT model is particularly valuable because it incorporates the discrimination parameter a_j, which quantifies how effectively a test item differentiates between models of varying abilities.

3.2. Fitting an IRT Model

We use the `py-irt` python package developed by Lalor and Rodriguez (2023) to leverage variational inference for IRT parameter estimations. Aligning with configurations explored in prior work, we employ a vague prior of $\mathcal{N}(0, 1)$ for θ_i and b_j, and $\log a_i \sim \mathcal{N}(0, \sigma_a^2)$ where $\sigma_a = 0.25$, the lower-end of the range searched in Vania et al. (2021). To assess sensitivity, we experimented with different prior configurations for σ_a and found that item and model parameters estimated under our chosen setup exhibited a high correlation ($r_\theta = 0.99, r_b = 0.99, r_a = 0.98$) with those from alternative setups. This suggests that our prior selection yields stable parameter estimates. To ensure that the learned IRT fit is reliable we examine and report the AUC-ROC scores in Appendix A, finding a range from 0.87 to 0.92.

3.3. Data & Model Selection

We focused our experiments on three benchmarks that are used to assess mathematical ability: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and MathOdyssey (Fang et al., 2024). GSM8K is a prominent dataset for assessing mathematical ability composed entirely of grade school-level math word problems, while MATH focuses on competition-level math questions. MathOdyssey is a new dataset that has not been widely used at this time but contains problems with a diverse range of difficulties, created to provide insight into emerging capabilities of LLMs and distinguish between increasingly powerful models. We collect responses for the entire test set of GSM8K and MathOdyssey, and collect responses to 500 questions of the MATH dataset (the full MATH dataset is not publicly available). See Appendix B for more information on the benchmark datasets.

Our model selection prioritized models of different sizes (Appendix C), with the populations' number of parameters ranging from 0.5 billion with Qwen2.5-0.5B-Instruct to an estimated 1 trillion with GPT-4.

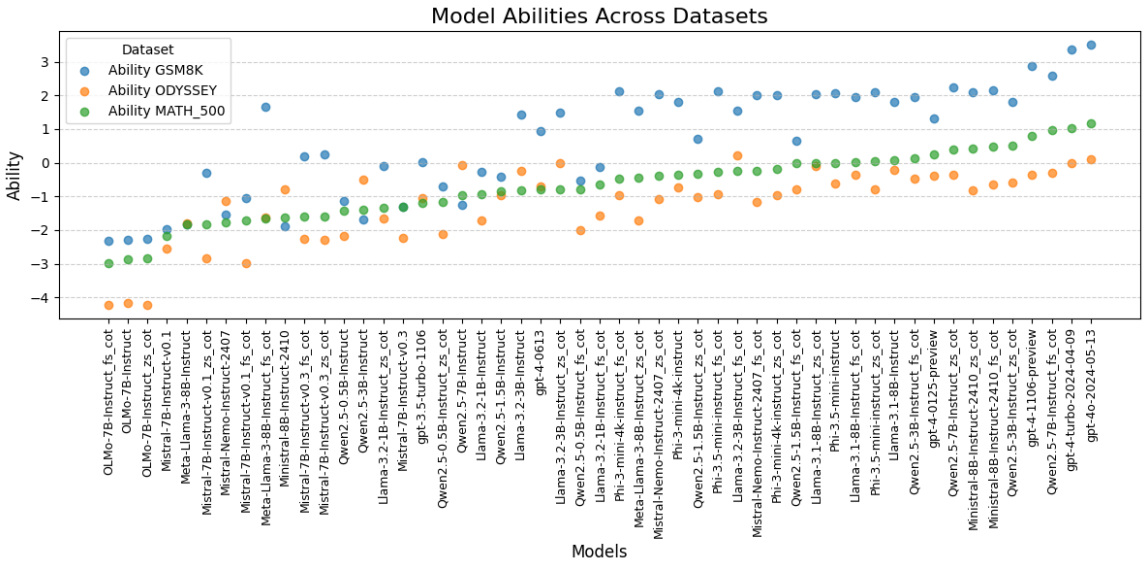
A good IRT fit requires a large population of test-takers. To maximize the population size we simulated three separate test-takers per model using different prompting techniques: zero-shot with no chain-of-thought (CoT), zero-shot with CoT, and few-shot with CoT. These prompting techniques have been shown to improve a model's reasoning skills (Wei et al., 2023). Our prompting strategies can be found in Appendix D.

4. Results and Analysis

As expected, the models perform extremely well on the GSM8K benchmark, while performance on MATH and MathOdyssey is poorer. This discrepancy in performance is seen in

the model ability estimation by our 2-PL IRT models for the GSM8K, MATH, and MathOdyssey dataset, shown in Figure 1. For almost all models, the estimated ability on GSM8K is higher than that of MATH and MathOdyssey.

Figure 1: Estimated Model Abilities for 2-PL IRT Model on GSM8k, MATH, and MathOdyssey



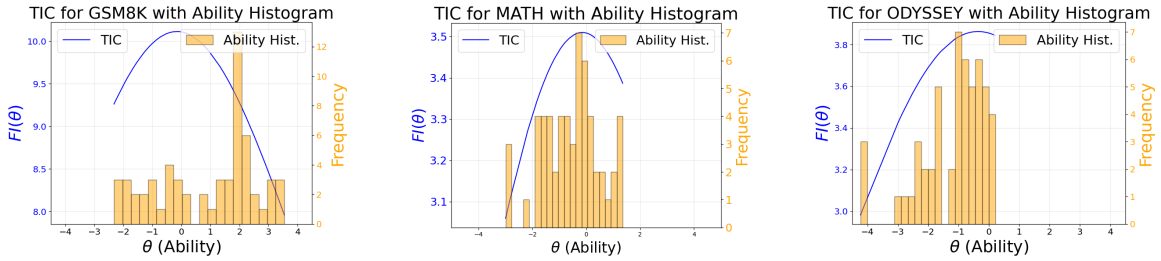
4.1. Robustness of Benchmark Ability Estimation

To evaluate how much information a given item provides we use the Item Information Function (IIF) based on item discrimination and the probabilities of a correct or incorrect response. Equation (2) calculates the IIF for a 2-PL IRT model where $P_j(\theta, b_j, a_j)$ is the probability a model of ability θ will correctly answer item j and $Q_j(\theta, b_j, a_j)$ is the probability of an incorrect response. Because our analysis is more concerned with evaluating how useful a benchmark is overall, we calculate Equation (3), the Test Information Curve (TIC), which is the sum of Equation (2) over all items. Intuitively, the TIC tells us how informative a test is given the ability of a test-taker.

$$I_j(\theta, b_j, a_j) = a_j^2 P_j(\theta, b_j, a_j) Q_j(\theta, b_j, a_j) \tag{2}$$

$$T(\theta, \mathbf{b}, \mathbf{a}) = \sum_j I_j(\theta, b_j, a_j) \tag{3}$$

The results from our IRT analysis confirm our hypothesis that GSM8K overall does not give us much information about the ability of state-of-the-art models. As seen in Figure 2(*a*), models with $\theta = 0$ achieve maximum test information. However, the learned θ values of the different models are skewed much higher, with a significant population of models having $\theta \geq 1$. The rest of the TICs in Figure 2 display evidence that the more difficult benchmarks



(a) GSM8K (b) MATH (c) MathOdyssey

Figure 2: Test Information Curves for GSM8K, MATH, and MathOdyssey benchmarks alongside the respective model ability distributions for both IRT fits. The TIC tells us how much information the test provides about a test-taker with a given ability. Because model abilities are fit based on the response patterns for each benchmark, ability range varies with benchmark difficulty. We find that MATH is best suited for the models in our setup.

of MATH and MathOdyssey are better suited for the current suite of LLMs. However, as model abilities improve, MATH and MathOdyssey may become ill-suited to differentiate between models. Using IRT provides us with a systematic way of analyzing the effectiveness of benchmarks, and can also help us select the most difficult and discriminating questions out of a dataset.

4.2. Rethinking Rankings

Leaderboards are commonly used to compare the performance of models and make conclusions on their relative ability based on overall accuracy on a specific benchmark dataset. Building on Rodriguez et al. (2021), we explore the stability of such rankings. We split the dataset questions into three groups - high, medium, and low - based on their estimated discrimination parameter, which measures the ability of a question to discriminate between model abilities. We then compare model rankings derived from accuracy on each subset to those based on overall accuracy and the estimated ability θ_i.

Figure 3 shows how the rankings of the top 20 models by ability on each benchmark shift when using different metrics or subsets of questions. The rankings remain relatively stable between the overall accuracy and the estimated abilities (θ), but fluctuate when based on performance on high discrimination questions and low discrimination questions. For a benchmark to confidently estimate relative model ability, model rankings should remain stable across different question subsets of the same benchmarks.

For GSM8K, we observe stability in the very top few models, however, we see significant fluctuation for several models when ranked on the highest discrimination subset. We see extreme fluctuations when ranking by low discrimination questions, indicating their inability to reliably distinguish between model abilities. We hypothesize that the low discrimination questions capture noise rather than ability. However, for MATH, we see that rankings are relatively stable between ability, overall accuracy, and accuracy on the high discrimination

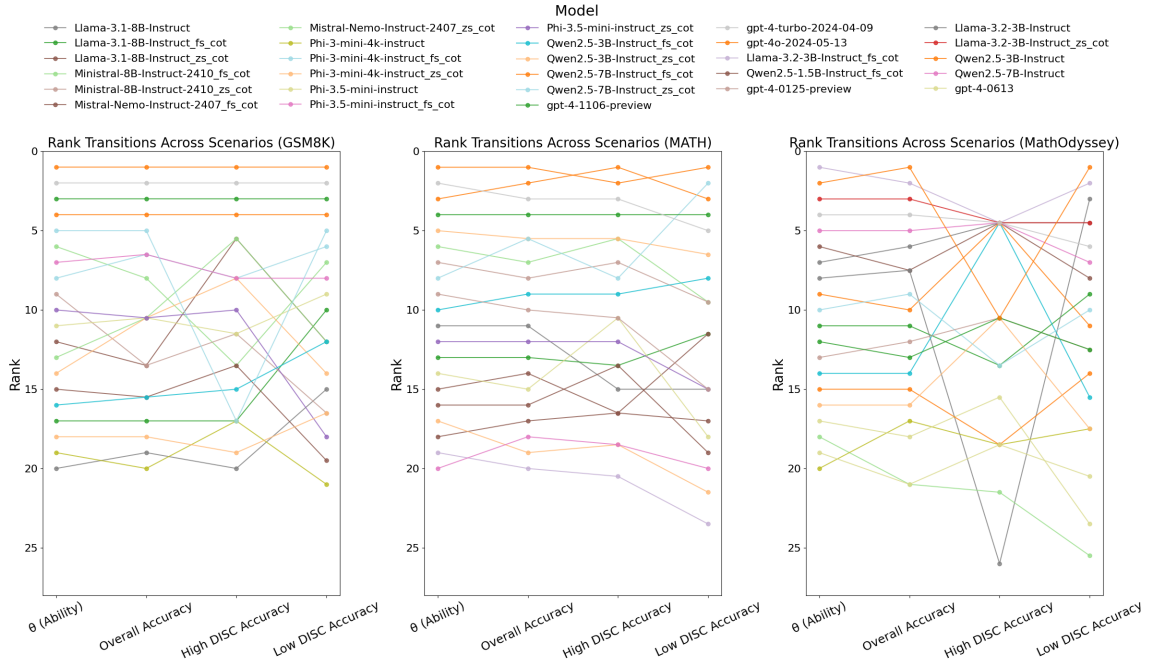


Figure 3: Changes in model rankings when using different metrics and benchmark subsets. This chart only shows the change in rankings for the top 20 models based on estimated model ability on each benchmark. We see that model rankings change when using highly discriminative subsets in comparison to overall benchmark accuracy or estimated model ability (θ_i).

subset. We hypothesize that this is due to the goodness-of-fit of the TIC with the MATH benchmark, shown in Figure 2(*b*), meaning that the MATH benchmark is well-suited for our model suite. We again see fluctuation in ranking on low discrimination questions. Finally, for MathOdyssey, we observe that several of the top performing models collapse to the same rank when compared based on the high discrimination subset, indicating that differences in ranking based on overall accuracy may be spurious and unreliable.

The instability in rankings across benchmarks when comparing model rank based on overall benchmark accuracy, estimated ability θ, and benchmark subsets accuracy makes it challenging for practitioners to effectively conclude that one model is better than another at the latent ability we are measuring (ability to do mathematical reasoning). Our framework grounds ability estimation and ranking in highly discriminative questions, which are better able to discriminate between model abilities in comparison to overall accuracy.

5. Discussion

Our findings show that current math benchmarks fall short in estimating model abilities and distinguishing between frontier models. Given that AIEd systems for math instruction depend heavily on choosing models with high mathematical reasoning proficiency, develop-

ers need robust and reliable methods for ranking model performance in this domain. Small fluctuations in accuracy on math benchmarks between SOTA models are not sufficient. Additionally, when designing benchmarks for future AIEd systems, stakeholders must thoroughly understand model capabilities to avoid adverse impacts on students and teachers (Holstein and Doroudi, 2021). Using IRT, developers can create benchmarks that more effectively differentiate between model abilities by selecting test items that have high discrimination parameters, providing more reliable performance rankings. We propose using IRT to curate subsets of benchmarks that include items based on their information value for models within a specific ability range, mirroring how educational tests are designed using IRT. IRT can also be used to evaluate new questions, measuring their discrimination and difficulty parameter to assess whether they should be added to an existing benchmark. This shows promise in both improving the discrimination of benchmarks and increasing their difficulty over time. Future research should validate this approach by curating subsets of benchmarks using IRT and testing their reliability and usefulness for downstream tasks.

5.1. Limitations

While we tested the reliability of our IRT fit, there are a few limitations of the data this fit relies on. First, we ran each model on each item only once. Future work should prompt each model several times to account for prompt sensitivity. Our method for checking the models' responses also relied on either heuristics or evaluation by GPT-4o, introducing the risk that a model's response may have been mischaracterized. Lastly, except for OpenAI models, we limited our largest model to 14B parameters due to compute constraints, but our IRT fit may have been stronger had we used a more even range of model sizes.

6. Conclusion

Given recent research calling into question the validity of popular math benchmarks for LLMs, we use IRT to evaluate the effectiveness of these benchmarks at discriminating between several SOTA models. We find that **(a)** GSM8K misses the mark given the current landscape, providing limited information for the current range of SOTA abilities and **(b)** model rankings based on overall accuracy can be unreliable across datasets. Our method also serves as a framework for evaluating the appropriateness of a benchmark for a range of model abilities and better estimating relative model ranks using highly discriminative questions. Our findings point to the promise of using IRT in the future development and evaluation of benchmarks as model abilities improve.

References

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt,

Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, 2024. URL https://arxiv.org/abs/2404.09932.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, April 2023. doi: 10.1126/science.adf6369.

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Item Response Theory ... A Statistical Framework for Educational and Psychological Measurement. 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Michael C. Edwards. An Introduction to Item Response Theory Using the Need for Cognition Scale. *Social and Personality Psychology Compass*, 3(4):507–529, 2009. doi: 10.1111/j.1751-9004.2009.00194.x. URL https://doi.org/10.1111/j.1751-9004.2009.00194.x.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Psychology Press, New York, 1 edition, 2000. ISBN 9781410605269. doi: 10.4324/9781410605269. URL https://psycnet.apa.org/record/2000-03918-000.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. MathOdyssey: Benchmarking Mathematical Problem-Solving Skills in Large Language Models Using Odyssey Math Data, 2024. URL https://arxiv.org/abs/2406.18321.

Ronald K. Hambleton and Hariharan Swaminathan. *Item Response Theory: Principles and Applications*. Springer Dordrecht, Dordrecht, 1 edition, 1985. ISBN 978-0-89838-065-1. doi: 10.1007/978-94-017-1988-9. URL https://doi.org/10.1007/978-94-017-1988-9. Springer Science+Business Media New York.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Kenneth Holstein and Shayan Doroudi. Equity and Artificial Intelligence in Education: Will "AIEd" Amplify or Alleviate Inequities in Education?, 2021. URL https://arxiv.org/abs/2104.12920.

John P. Lalor, Hao Wu, and Hong Yu. Building an Evaluation Scale using Item Response Theory, September 2016. URL http://arxiv.org/abs/1605.08889. arXiv:1605.08889 [cs].

John P. Lalor, Hao Wu, and Hong Yu. Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds, August 2019. URL http://arxiv.org/abs/1908.11421. arXiv:1908.11421 [cs].

John Patrick Lalor and Pedro Rodriguez. py-irt: A Scalable Item Response Theory Library for Python. *INFORMS Journal on Computing*, 35(1):5–13, January 2023. ISSN 1526-5528. doi: 10.1287/ijoc.2022.1250. URL http://dx.doi.org/10.1287/ijoc.2022.1250.

Yucheng Li. Estimating Contamination via Perplexity: Quantifying Memorisation in Language Model Evaluation, 2023. URL https://arxiv.org/abs/2309.10677.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Wim J. Linden and Ronald K. Hambleton, editors. *Handbook of Modern Item Response Theory*. Springer New York, NY, New York, NY, 1 edition, 1997. ISBN 978-0-387-94661-0. doi: 10.1007/978-1-4757-2691-6. URL https://doi.org/10.1007/978-1-4757-2691-6. Springer Science+Business Media New York.

Pepper Miller and Kristen DiCerbo. LLM-Based Math Tutoring: Challenges and Dataset, 07 2024.

Melvin R. Novick. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1):1–18, February 1966. ISSN 0022-2496. doi: 10.1016/0022-2496(66)90002-2. URL https://www.sciencedirect.com/science/article/pii/0022249666900022.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Rahul Ramesh, Ekdeep Singh Lubana, Mikail Khona, Robert P. Dick, and Hidenori Tanaka. (compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks), 2024. URL https://arxiv.org/abs/2311.12997.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis 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, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346.

Clara Vania, Phu Mon Htut, William Huang, Dhara A. Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing Test Sets with Item Response Theory. *CoRR*, abs/2106.00840, 2021. URL https://arxiv.org/abs/2106.00840.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise, 2024. URL https://arxiv.org/abs/2410.03017.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023. URL https://arxiv.org/abs/2201.11903.

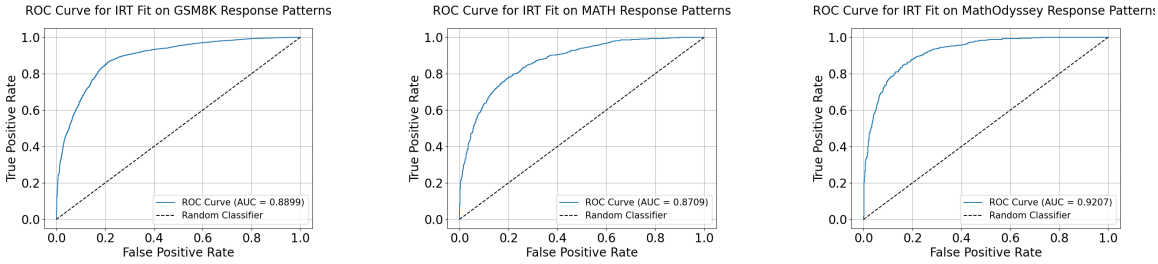
Mike Wu, Richard L. Davis, Benjamin W. Domingue, Chris Piech, and Noah Goodman. Variational Item Response Theory: Fast, Accurate, and Expressive, March 2020. URL http://arxiv.org/abs/2002.00276. arXiv:2002.00276 [cs, stat].

Tianlong Xu, Yi-Fan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. AI-Driven Virtual Teacher for Enhanced Educational Efficiency: Leveraging Large Pretrain Models for Autonomous Error Analysis and Correction, 2024. URL https://arxiv.org/abs/2409.09403.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-Mix: a Flexible and Expandable Family of Evaluations for AI models, 2023. URL https://arxiv.org/abs/2310.17567.

Appendix A. IRT Model Fit

Figure 4 shows the ROC curves for our IRT fit on our three benchmarks of interest. We find AUC-ROC scores ranging from 0.87 to 0.92, indicating a reliable IRT fit.



(a) GSM8K (b) MATH (c) MathOdyssey

Figure 4: ROC Curves for the IRT Fit on (a) GSM8K response patterns with AUC-ROC Score = 0.8899 (b) MATH response patterns with AUC-ROC Score = 0.8709 and (c) MathOdyssey response patterns with AUC-ROC Score = 0.9207.

Appendix B. Benchmark Information

The table below provides a brief overview of the three benchmarks.

Dataset	Year	Description	Sample Size
GSM8k	2021	Middle-school level math word problems	1318
MATH	2021	Competition-level math problems. We use MATH-500, a subset of the original MATH dataset created by Lightman et al. (2023).	500
MathOdyssey	2024	High School, University, & Olympiad-level math problems	387

Appendix C. Models

The table below details information about the models used as learners in our work. The suffix '_fs_cot' indicates that the few-shot with chain-of-thought (CoT) version of the prompt was used, whereas _zs_cot' indicates that the zero-shot CoT version was used. A '-' indicates that we could not find this information published, or that it is unknown to the public.

Index	Learner Name	Num Params	Context Length	Family	Release Date
1	Qwen2.5-0.5B-Instruct	0.49	32768	Qwen	09_2024

Index	Learner Name	Num Params	Context Length	Family	Release Date
2	Qwen2.5-0.5B-Instruct_zs_cot	0.49	32768	Qwen	09_2024
3	Qwen2.5-0.5B-Instruct_fs_cot	0.49	32768	Qwen	09_2024
4	Qwen2.5-1.5B-Instruct	1.54	32768	Qwen	09_2024
5	Qwen2.5-1.5B-Instruct_zs_cot	1.54	32768	Qwen	09_2024
6	Qwen2.5-1.5B-Instruct_fs_cot	1.54	32768	Qwen	09_2024
7	Qwen2.5-3B-Instruct	3.09	32768	Qwen	09_2024
8	Qwen2.5-3B-Instruct_zs_cot	3.09	32768	Qwen	09_2024
9	Qwen2.5-3B-Instruct_fs_cot	3.09	32768	Qwen	09_2024
10	Qwen2.5-7B-Instruct	7.61	131072	Qwen	09_2024
11	Qwen2.5-7B-Instruct_zs_cot	7.61	131072	Qwen	09_2024
12	Qwen2.5-7B-Instruct_fs_cot	7.61	131072	Qwen	09_2024
13	Qwen2.5-14B-Instruct	14.7	131072	Qwen	09_2024
14	Qwen2.5-14B-Instruct_zs_cot	14.7	131072	Qwen	09_2024
15	Qwen2.5-14B-Instruct_fs_cot	14.7	131072	Qwen	09_2024
16	OLMo-7B-Instruct	6.89	2048	OLMo	02_2024
17	OLMo-7B-Instruct_zs_cot	6.89	2048	OLMo	02_2024
18	OLMo-7B-Instruct_fs_cot	6.89	2048	OLMo	02_2024
19	OLMoE-1B-7B-0924-Instruct	6.92	-	OLMo	09_2024
20	Llama-3.2-1B-Instruct	1.24	128000	Llama	09_2024
21	Llama-3.2-1B-Instruct_zs_cot	1.24	128000	Llama	09_2024
22	Llama-3.2-1B-Instruct_fs_cot	1.24	128000	Llama	09_2024
23	Llama-3.2-3B-Instruct	3.21	128000	Llama	09_2024
24	Llama-3.2-3B-Instruct_zs_cot	3.21	128000	Llama	09_2024
25	Llama-3.2-3B-Instruct_fs_cot	3.21	128000	Llama	09_2024
26	Meta-Llama-3-8B-Instruct	8.03	8000	Llama	04_2024
27	Meta-Llama-3-8B-Instruct_fs_cot	8.03	8000	Llama	04_2024
28	Meta-Llama-3-8B-Instruct_zs_cot	8.03	8000	Llama	04_2024
29	Llama-3.1-8B-Instruct	8.03	128000	Llama	07_2024
30	Llama-3.1-8B-Instruct_zs_cot	8.03	128000	Llama	07_2024
31	Llama-3.1-8B-Instruct_fs_cot	8.03	128000	Llama	07_2024
32	Phi-3.5-mini-instruct	3.82	128000	Phi3	08_2024
33	Phi-3.5-mini-instruct_zs_cot	3.82	128000	Phi3	08_2024

Index	Learner Name	Num Params	Context Length	Family	Release Date
34	Phi-3.5-mini-instruct_fs_cot	3.82	128000	Phi3	08_2024
35	Phi-3-mini-4k-instruct	3.82	4000	Phi3	06_2024
36	Phi-3-mini-4k-instruct_zs_cot	3.82	4000	Phi3	06_2024
37	Phi-3-mini-4k-instruct_fs_cot	3.82	4000	Phi3	06_2024
38	Phi-3-medium-4k-instruct	14	4000	Phi3	06_2024
39	Phi-3-medium-4k-instruct_zs_cot	14	4000	Phi3	06_2024
40	Phi-3-medium-4k-instruct_fs_cot	14	4000	Phi3	06_2024
41	Mistral-7B-Instruct-v0.1	7.24	-	Mistral	09_2023
42	Mistral-7B-Instruct-v0.1_zs_cot	7.24	-	Mistral	09_2023
43	Mistral-7B-Instruct-v0.1_fs_cot	7.24	-	Mistral	09_2023
44	Mistral-7B-Instruct-v0.3	7.25	-	Mistral	05_2024
45	Mistral-7B-Instruct-v0.3_zs_cot	7.25	-	Mistral	05_2024
46	Mistral-7B-Instruct-v0.3_fs_cot	7.25	-	Mistral	05_2024
47	Ministral-8B-Instruct-2410	8.19	-	Mistral	10_2024
48	Ministral-8B-Instruct-2410_zs_cot	8.19	-	Mistral	10_2024
49	Ministral-8B-Instruct-2410_fs_cot	8.19	-	Mistral	10_2024
50	Mistral-Nemo-Instruct-2407	8.19	-	Mistral	10_2024
51	Mistral-Nemo-Instruct-2407_zs_cot	12.2	-	Mistral	07_2024
52	Mistral-Nemo-Instruct-2407_fs_cot	12.2	-	Mistral	07_2024
53	gpt-3.5-turbo-1106	-	16385	OpenAI	-
54	gpt-3.5-turbo	-	16385	OpenAI	-
55	gpt-4-0125-preview	-	128000	OpenAI	-
56	gpt-4-0613	-	8192	OpenAI	06_2023
57	gpt-4-1106-preview	-	128000	OpenAI	-
58	gpt-4-turbo-2024-04-09	-	128000	OpenAI	09_2024
59	gpt-4o-2024-05-13	-	128000	OpenAI	05_2024
60	gpt-4o-2024-08-06	-	128000	OpenAI	08_2024
61	gpt-4o-mini-2024-07-18	-	128000	OpenAI	07_2024
62	gpt-4o	-	128000	OpenAI	08_2024

Appendix D. Prompting Strategies

The following prompts (Figure 5(a), Figure 5(b), Figure 5(c)) were used across all models. For few-shot CoT prompting we used the 3 questions selected from the dataset of interest. We confirmed that these questions were not in the samples selected for evaluation.

Appendix E. Extracting Model Answers

For GSM8K and MATH, we extract the final number from the model's response, using this as the models "final answer." We prompted models to use a specific format when answering questions, but the small size and ability of some models made response structure inconsistent, forcing us to use this heuristic. We validated this mechanism for scraping answers by manually inspecting a subset of the responses. We then used regular expression matching to check model answers.

Since MathOdyssey is a significantly more challenging benchmark and the questions are represented in Latex rather than plain-text formatting, the developers of the benchmark use LLM-supported answer checking (Fang et al., 2024), so we also check model responses using OpenAI GPT-4o.

```
Base Prompt:

You are a helpful assistant. Answer the following
question accurately and concisely. Label your
final answer with ### Answer: [final answer].

Question: What is 2+2? ### Answer: 4. Question:
What is 5 * 10? ### Answer: 50",
```

(*a*) Base prompt used

```
Zero-shot Chain-of-Thought Prompt:

You are a helpful assistant. Answer the following
question accurately and concisely. Let's think
step by step. Label your final answer with ###
Answer: [final answer].

Question: What is 2+2? ### Answer: 4. Question:
What is 5 * 10? ### Answer: 50",
```

(*b*) Zero-Shot CoT prompt used

```
Few-shot Chain-of-Thought Prompt:

You are a helpful assistant. Answer the following
question accurately and concisely. Let's think
step by step and check your work as you go. Here
are some examples: Question: {EXAMPLE_QUESTION_1}
Reasoning:{REASONING_1} ### Answer: {ANSWER_1}
Question: {EXAMPLE_QUESTION_2}
Reasoning:{REASONING_2} ### Answer: {ANSWER_2}
Question: {EXAMPLE_QUESTION_3}
Reasoning:{REASONING_3} ### Answer: {ANSWER_3}

Be sure to label your final answer with ###
Answer: [final answer].
```

(*c*) Few-Shot CoT prompt used. Examples were substituted in based on the dataset and kept consistent across all models.

Figure 5: Prompts used to collect model response patterns. Each model was prompted once for each question using three techniques.