How Hard is Math? Using Quantitative Metrics to Measure LLM Alignment to Human Intuitions of Difficulty

Micah Helzerman, Steven R. Wilson, and Cam McLeman
University of Michigan-Flint
{mhelzerm, steverw, mclemanc}@umich.edu

Abstract

Modern LLMs have demonstrated advanced reasoning skills, including the ability to solve Olympiad-level mathematics problems. While solving more and more difficult problems is a hallmark of LLM progress, less attention has been placed on how "difficulty" is operationalized in the context of LLM problem solving tasks. This is particularly relevant in educational contexts where teachers or students may ask LLMs for "easy" or "hard" questions. In this paper, we explore various quantitative measurements from LLM-generated solutions and evaluate their inter-correlations, as well as their correlation to human-annotated difficulty scores. We find moderate correlations between metrics using log probabilities and output lengths, including some that are more strongly correlated to difficulty than LLM accuracy. We also train ModernBERT to predict difficulty scores, leading to reasonable accuracy within a given benchmark, but decreased performance when generalizing to other math benchmarks. Finally, to explore connections between difficulty scores and human performance, we collect problems, human solutions, and human performance data from the Putnam competition. We find poor alignment between LLM metrics and human-assigned difficulty scores, despite strong correlations between those scores and human performance on the problems.

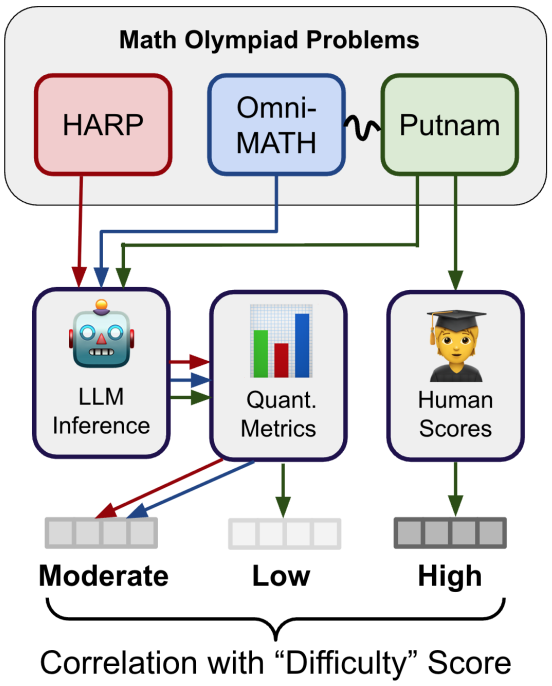


Figure 1: When LLMs solve Olympiad-level mathematics problems, quantitative metrics can be extracted which have moderate correlations (~0.35-0.40) to expert-annotated difficulty scores. These correlations become weak (< 0.10) when looking at very difficult problems, i.e. from the Putnam Competition, despite real human scores achieving the highest correlation to difficulty scores (0.80).

1 Introduction

LLMs have grown increasingly powerful in their level of detail and accuracy when solving math problems, demonstrating advanced reasoning capabilities with chain-of-thought (Wei et al., 2022; Kojima et al., 2022). This has led to researchers developing more difficult mathematics benchmarks, such as HARP and Omni-MATH, pushing LLMs to their limit (Yue et al., 2024; Gao et al., 2025). Often overlooked is the fact that this requires some definition of "difficulty"[1], which could have different meanings in the context of mathematics: the total effort and time required to solve a problem; the requisite skills and problem-solving techniques, often based on experience; the uncertainty felt when solving a problem. For modern mathematics benchmarks, which draw problems from Olympiad-level

[1]While we note the broad range of definitions for difficulty in mathematics, and the knowledge gap this creates in literature, for the purposes of this paper we will focus on the human-annotated difficulty score common to modern mathematics benchmarks.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 968–981

competitions, difficulty is typically annotated by subject matter experts using a scale directly related to the recommended/required age of participants (e.g., AMC-8, AMC-10, and AMC-12, designed for middle school and high school students below grades 8, 10, and 12, respectively). While it is standard to measure the accuracy of an LLM as difficulty increases, we explore the extent to which quantitative metrics—the number of generated output tokens (effort), whether the problem includes Asymptote language (geometric spatial skills), and log probabilities of generated solutions (uncertainty)—of LLM solutions relate to human-assigned difficulty ratings. The amount of effort, mathematical skills used, and uncertainty felt when a student is solving a problem, can be used to measure the difficulty of a math problem. Using Pearson correlation, we find that several metrics moderately align to human difficulty ratings—even more strongly than LLM solution accuracy—and show that they have some predictive power for a given problem's difficulty ratings, supporting recent findings (Plaut et al., 2025; Luo et al., 2025).

However, these metrics do not explain all of the variance; notably, it is missing token- and sentence-level semantics. To help bridge this gap, we fine-tune ModernBERT (Warner et al., 2024) to classify human difficulty ratings using a problem's text and a provided human solution. We find that for problems from HARP (6 difficulty levels) and Omni-MATH (9 difficulty levels), ModernBERT can be trained to achieve 42-57% accuracy (~90% within-1 accuracy) when tested on Olympiad-level problems within their datasets, which decreases significantly (29-37% accuracy) when tested on Olympiad-level problems in the opposite benchmark, suggesting that ModernBERT is unable to learn a generalizable representation of difficulty from semantics.

We also explore a subset of the hardest problems in Omni-MATH from the Putnam Competition, enabling us to make a direct comparison between how well human competitors scored [2] compared to LLMs. We find that, while real human scores correlate strongly to the difficulty ratings, the correlation between LLM accuracy and difficulty ratings disappears (Figure 1). Our contributions include (1) a comprehensive evaluation of the correlation between quantitative metrics from LLMs, human-annotated difficulty scores, and human performance, and (2) fine-tuning ModernBERT on ordinal regression of difficulty level given problem text and solution, utilizing two modern mathematics benchmarks and an undergraduate competition with human scores for both contributions.

[2]We use https://kskedlaya.org/putnam-archive/ to collect real human scores.

2 Related Work

2.1 Mathematics Benchmarks

To date, there have been dozens of mathematics benchmarks developed to test the mathematical capabilities of large language models. MATH was the first large dataset of 12,500 competition-level problems written in natural language, with state-of-the-art models (at the time) scoring between 3.0% and 6.9% (Hendrycks et al., 2021). While problems from MATH are considered to be difficult, coming from competitions taken by "the best young mathematical talent in the United States" (Hendrycks et al., 2021), early LLMs still struggled with mathematical benchmarks that were based on grade-school curriculum. A second example, GSM8k consists of 8,500 grade-school math problems that require knowing how to apply the four basic arithmetic operations to solve, yet GPT-3 175B could only correctly solve around 55% of the problems given 100 attempts each (Cobbe et al., 2021). Modern LLMs use chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) and self-consistency among multiple generated solutions (Wang et al., 2023), along with more parameters and larger training sets, to achieve higher accuracies. Google's Gemini 1.5 Pro achieved 67.7% accuracy on MATH and 90.8% accuracy on GSM8k (Team et al., 2024), and OpenAI's o1 reasoning model had a 94.8% accuracy on MATH with only one attempt per solution (OpenAI, 2024).

With these improvements to LLM accuracy for the MATH and GSM8k benchmarks, researchers have introduced new benchmarks: HARP collects the most difficult problems from prestigious mathematical competitions like A(J)HSME, AMC, AIME, and USA(J)MO (Yue et al., 2024), and Omni-MATH contains problems from all prestigious mathematics competitions as listed on the AoPS (Art of Problem Solving) webpage (Gao et al., 2025). Top performing models (at the time of benchmark publication) struggle with problems from HARP (58.1% for Gemini 1.5 Pro, and 75.9% for o1-mini) and Omni-MATH (60.54% for o1-mini).

Notable with the introduction of the HARP and Omni-MATH benchmarks is the attention toward assigning problems a difficulty score based on human judgments. Both papers collect this score from subject matter experts of mathematics competitions (AoPS), who define a 10-point difficulty scale to compare between the different mathematical competitions (Art of Problem Solving, 2025). While we do not believe this can be used as a thorough metric covering the multiple facets of human intuitions of difficulty, such as problem structure or the skills necessary to solve each problem (Lucy et al., 2024), we use it in this work in order to directly engage with findings in prior work that share this definition. We leave the systematic formation of more thorough definitions of difficulty for future work.

2.2 Using `logprobs` to Measure LLM Confidence

A key metric used to evaluate the performance of large language models is perplexity, a score measuring how "surprised" a model is when predicting tokens (Jurafsky and Martin, 2025). Perplexity is the inverse of token-level log probabilities, meaning a model with low perplexity will generate tokens with higher log probabilities.

Prior work in uncertainty quantification (UQ) shows that token probabilities can be an effective measure of model uncertainty (Malinin and Gales, 2021), and can be improved by weighting each token based on its relevancy to the generation (Duan et al., 2024a). Plaut et al. (2025) find that these log probabilities can be used to predict an LLM's accuracy when solving multiple-choice questions, even across models of different sizes and architectures. Orgad et al. (2025) explores various metrics of log probabilities, including taking the minimum, average, and maximum values across a prompt, as well as looking at the log probabilities of answer tokens.

2.3 Using BERT Family Models for Mathematics

BERT has been used in different mathematics contexts, including mathematics understanding (Peng et al., 2021; Shen et al., 2023), mathematics reasoning (Piękos et al., 2021) as well as subject and difficulty classification (Lao and Lei, 2023; Duan et al., 2024b). ModernBERT is a modernized version of BERT, with improvements to architecture that allow for a larger context window (8192 tokens) and faster inference (Warner et al., 2024).

3 Experimental Setup: LLM Metrics

We selected LLMs that were included in the original HARP and/or Omni-MATH benchmarks, including 5 local models and 1 API-based model: `gpt-4o-mini`, `gemma-3-27b-it`, `Llama-3.1-8B-Instruct`, `Mathstral-7B-v0.1`, `NuminaMath-7B-CoT`, and `Qwen2.5-Math-7B-Instruct`. Following the work of Yue et al. (2024) and Gao et al. (2025), we set each model's temperature parameter to 0 and top_p parameter to 1 to reduce variations in model responses. For each model, we use the recommended system prompt, then prompt on a random sample of 1000 questions from each of the two benchmarks, and record their output tokens and corresponding log probabilities.

Additionally, we collected 360 Putnam problems, around 100 of which are included in Omni-MATH. The purpose of including a separate dataset for only Putnam problems is that we can include real human scores, allowing us to compare the accuracy of human participants to our LLM metrics. These problems are at difficulty levels 7, 8, and 9 (according to AoPS), making them some of our hardest problems. We prompted the same LLMs on all 360 problems, again recording the output tokens and corresponding log probabilities.

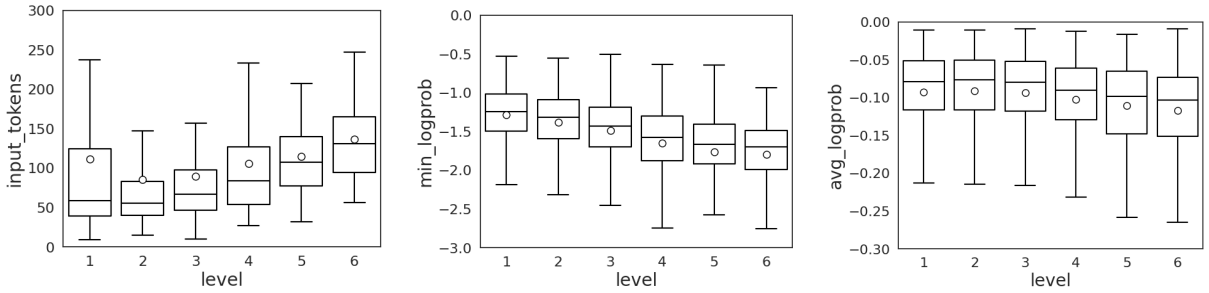
3.1 Evaluating Solution Correctness: HARP vs. Omni-MATH

Both benchmarks rely on separate systems for automatic evaluation of solution correctness. Yue et al. (2024) use a sympy-based parser and answer checker for HARP, whereas Gao et al. (2025) introduce a GPT-4o-based model, Omni-Judge, which has 86% consistency with human graders. We decided to use Omni-Judge, as it can more accurately grade natural language variations in answers (e.g. "`$4.00\\text{ dollars}$`" vs "`$\\textdollar4$`)" and therefore reduce false negatives.

3.2 Structure of our data

The structure of our data includes the following:

- Full problem text in natural language, as well as the full LLM solution
- Problem difficulty (`level`) based on human expert judgments



(a) Distribution of `input_tokens`. (b) Distribution of `min_logprob`. (c) Distribution of `avg_logprob`.

Figure 2: Distributions of model-related statistics grouped by difficulty level for HARP problems.

- Number of tokens in the math problem (`input_tokens`) and generated solution (`output_tokens`)
- Whether the LLM solution reached the correct answer (`is_correct`), judged using Gao et al.'s (2025) LLM-based answer checker (Omni-Judge)
- The log probabilities of each selected token in the generation (`top_logprobs`)
- Whether the problem uses Asymptote, a vector graphics language (`has_asy_problem`)

We choose to include `has_asy_problem` in our analysis, as Luo et al. (2025) have shown that LLMs struggle more with accurately answering these types of problems. We test to see whether `has_asy_problem` could be used as a potential signal for an LLM's notion of difficulty.

Further, we compute two additional metrics based on token log probabilities:

- `min_logprob`, which is calculated per LLM solution by taking the log probability from the token with the smallest log probability.
- `avg_logprob`, which is calculated per LLM solution by taking the average of the log probabilities of all tokens.

We also investigate these metrics for the answer tokens, following the work of Orgad et al. (2025). With these metrics, we can explore the relation between log probabilities and human difficulty labels. We do not consider `max_logprob`, as for all generated solutions this would be approximately 0.

Additionally in the HARP paper, Yue et al. (2024) found that LLMs scale up the number of tokens in their generated solutions as the problem difficulty increases. They describe this behavior as consistent across all tested LLMs, as well as human-generated solutions, suggesting models could be biased from human solutions in their underlying training data. However, this may be confounded by the number of tokens in the problems themselves, something that is underexplored in prior literature. We test to see whether `input_tokens` is correlated to `output_tokens` in the HARP and Omni-MATH data. While we do not set out to control for problem or solution length to determine causality, as the former would require redesigning the benchmark and the latter may artificially reduce model performance, we explore the solution-level correlation between these variables. To further explore potential interactions between LLM metrics and difficulty level, we conduct a series of OLS regressions (see Appendix C), using ablation to control for metrics and to see how these metrics interact. We find that adjusted R^2 is low, especially for Omni-MATH, and find that `input_tokens` and `min_logprob` has stronger correlation for difficulty level in Omni-MATH than in HARP.

4 Experimental Results: LLM Metrics

We first present our results visually, and then dive deeper into more specific and numeric data.

Problem Length (i.e., number of input tokens) scales with problem difficulty (Figure 2a). **Log Probabilities** decrease with problem difficulty, although the effect is more noticeable for `min_logprob` (Figure 2b) than for `avg_logprob` (Figure 2c). Both box-plots show there is high variance and significant overlap across level, meaning it is not enough information to accurately differentiate between problems by difficulty. See Appendix A for other distributions grouped by level.

Correlations Heatmap (Figure 3) shows moderate correlations between the three variables `output_tokens`, `min_logprob`, and `level` for

HARP, but only between `min_logprob` and `level` for Omni-MATH. This aligns with the finding from Yue et al. (2024) that LLM solutions increase in length as difficulty increases, and our observation that log probabilities (in the form of `min_logprob`) decrease as difficulty increases. However, we note that for Omni-MATH, there is a slightly lower correlation between `is_correct` and `level` than between `min_logprob` and `level`, while the inverse is true for HARP. Additionally, there is a lower correlation between `output_tokens` and `level` in Omni-MATH than in HARP, and a higher correlation between `input_tokens` and `level` in Omni-MATH than in HARP. Prior work found that log probabilities of exact answer tokens may be used to predict truthfulness (Orgad et al., 2025), however we find that `min_logprob_boxed` has low correlation to both `level` and `is_correct`.

We find that, while patterns are different across HARP and Omni-MATH, some LLM metrics may be more closely aligned to human difficulty labels than LLM accuracy. While trends in Figure 2a and Yue et al. (2024) show that problem length and LLM solution length both increase with difficulty level on average, there is only a very weak correlation between `input_tokens` and `output_tokens` (0.10 for HARP and 0.15 for Omni-MATH).

When solving Putnam problems, many of these patterns disappear (Figure 4). This includes correlations between `output_tokens`, `min_logprob`, and `level` (all < 0.10). In contrast, `human_score` has very strong correlation with `level` (0.80). This suggests that, while the difficulty score aligns strongly to human accuracy, it has little to no alignment to LLM accuracy. See Appendix B for correlation heatmaps of each LLM tested.

When comparing the accuracy between humans and LLMs (Figure 5), we find that LLMs have consistency across the 3 levels, helping to explain the correlations with the Putnam problems, and illustrating the poor alignment. The experts who create these competitions expect student accuracy to decrease while problems become increasingly difficult, however this is not necessarily the case for LLMs. While the pattern may appear at the macro-scale, i.e. when considering an entire benchmark of more than 4,000 problems, LLMs can only truly interact with one problem at a time, let alone a single competition. By de-aggregating the HARP and Omni-MATH benchmarks, we are able to get a more clear understanding of how LLMs are interacting with these problems of varying difficulty, lending greater insights into how LLMs are used by students and teachers at the micro level. See Appendix D for individual performance of each LLM tested.

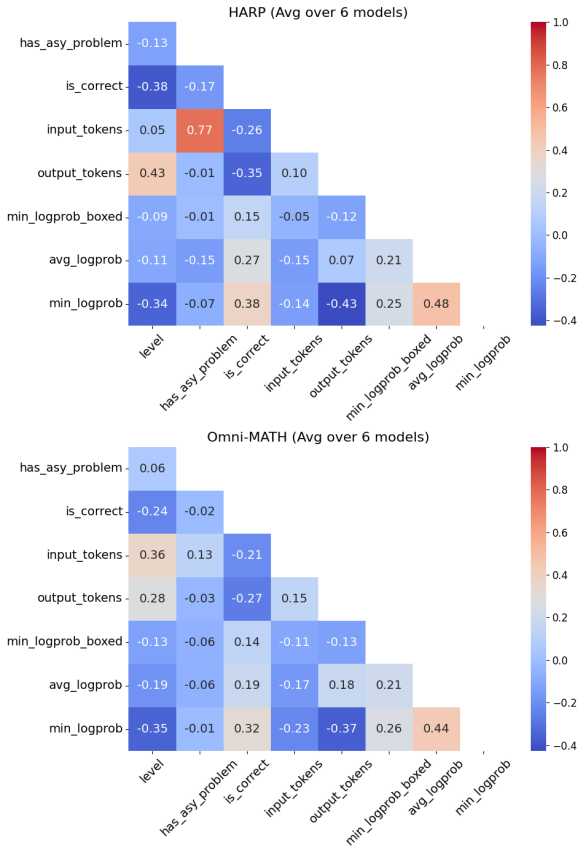


Figure 3: Pearson correlation heatmap of the variables from the HARP and Omni-MATH problems, averaged over the 6 models. Moderate correlations exist between `is_correct`, `level`, `min_logprob`, `input_tokens`, and `output_tokens`.

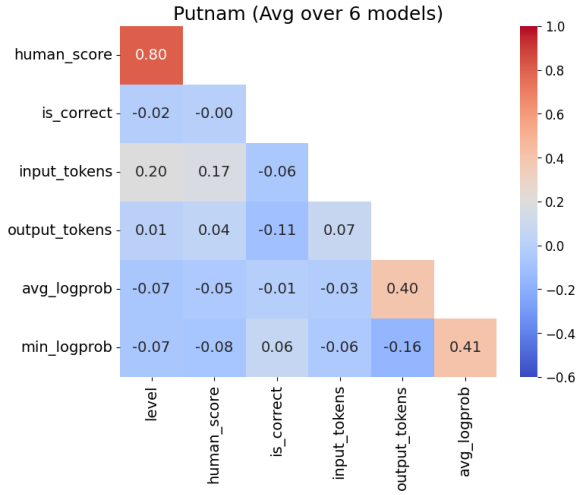


Figure 4: Pearson correlation heatmap of the variables from the Putnam problems, averaged over the 6 models.

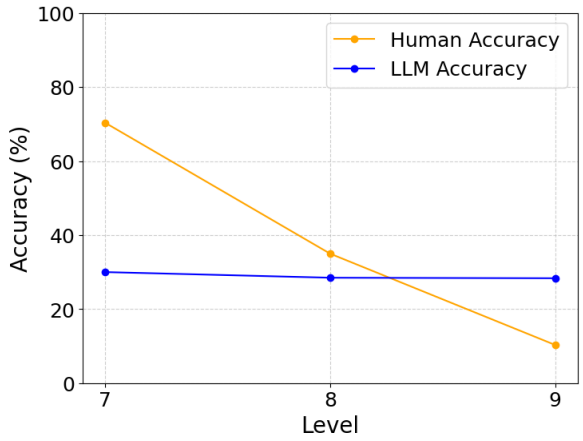


Figure 5: Accuracy for humans and LLMs when solving Putnam problems. While LLM accuracy remains constant as difficulty increases, human accuracy predictably decreases as problems get harder.

5 Experimental Setup: ModernBERT Difficulty Prediction

Instead of prompting LLMs to predict the difficulty score of a given problem, which can lead to hallucination or difficulty following the system prompt (especially for small models fine-tuned on step-by-step reasoning and solving mathematics problems), we explored using ModernBERT for classifying difficulty level of problems. To accommodate the math reasoning skills necessary to accurately classify a problem (Lucy et al., 2024), we incorporate text pairs using ModernBERT's tokenizer to combine both the problem and one of the human-written solutions (provided by the benchmark). Since the classification task is ordinal, we use an implementation of consistent rank logits (CORAL) (Cao et al., 2020) ordinal regression.

For all experiments using ModernBERT, we follow an 80-10-10 split for train/eval/testing. The train set is used to train a new model for 10 epochs ($2e-5$ learning rate, 0.01 weight decay). The evaluation set is used to evaluate the model at each epoch, from which we select the model with the highest F1 score to use on the test set. We record the accuracy, within-1 accuracy, and F1 scores.

We evaluate ModernBERT on HARP and Omni-MATH in three sub-experiments using (1) full HARP / Omni-MATH dataset, (2) 200 problems from each of six difficulty levels from ***one*** of HARP / Omni-MATH, and (3) 200 problems from each of six difficulty levels from ***both*** HARP / Omni-MATH. In these sub-experiments, we train ModernBERT on the given problems, and test it on both problems within its given test set and problems from the opposite benchmark test set.

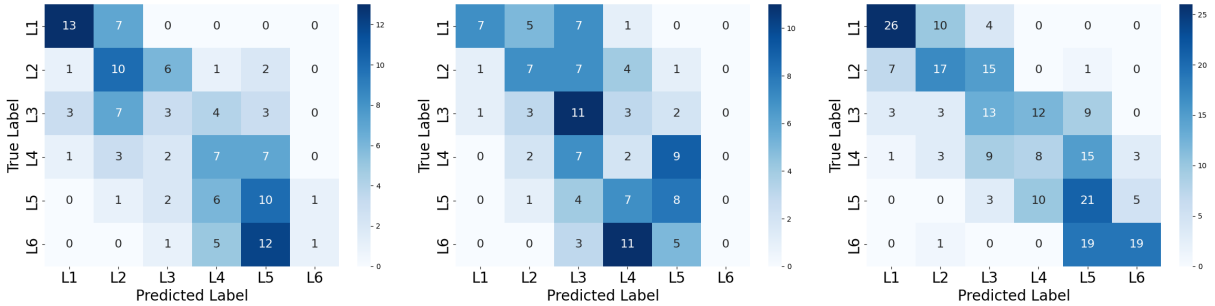
We also test ModernBERT on 360 Putnam problems in two sub-experiments, first using the full Putnam dataset for training, and second using Omni-MATH problems with difficulty levels in $\{7, 8, 9\}$, removing its Putnam problems, for training.

6 Experimental Results: ModernBERT Difficulty Prediction

Train Set	Test Set	MAE ↓	F_1^{mac} ↑
Full Dataset (Levels 1-6 HARP, 1-9 Omni-MATH)			
HARP	HARP	0.586	0.405
Omni-MATH	Omni-MATH	0.677	0.413
HARP	Omni-MATH	1.612	0.169
Omni-MATH	HARP	1.716	0.151
200 per Difficulty Level (Levels 1-6 Both Datasets)			
HARP	HARP	0.815	0.409
Omni-MATH	Omni-MATH	0.597	0.564
HARP	Omni-MATH	0.857	0.345
Omni-MATH	HARP	1.067	0.277
Mixed Model with Subsets (Levels 1-6 Both Datasets)			
Mixed	Mixed	0.696	0.443
Mixed	HARP	0.754	0.396
Mixed	Omni-MATH	0.635	0.479

Table 1: ModernBERT performance across datasets. MAE denotes the mean average error (↓ lower is better), F_1^{mac} denotes the Macro F_1 score (↑ higher is better).

For most models, we find exceptionally high within-1 accuracy, between 85-95%. The high within-1 accuracy suggests that ModernBERT captures a coarse notion of difficulty, even when exact prediction fails. It is more difficult for ModernBERT to predict difficulty level for HARP problems than Omni-MATH problems in all scenarios; this effect increases when training a model on problems from both datasets (Table 1). One potential cause is due to the differing ranges of competitions between the benchmarks: HARP contains problems from 4 unique benchmarks, whereas Omni-MATH contains problems from 27 unique benchmarks. Having fewer competitions may make it more difficult to label difficulty, as problems within the same competition could have similar semantics. The sharp decrease in accuracy and F1 scores when crossing the models on the full datasets is due to the differing range of levels, as the full Omni-MATH dataset contains problems at difficulties 1 through 9, and the full HARP dataset contains problems at difficulties 1 through 6. The mismatch results in the



(a) HARP model tested on Omni-MATH (36.97% accuracy, 0.857 mean average error, 34.48% F1 score).

(b) Omni-MATH model tested on HARP (29.41% accuracy, 1.067 mean average error, 27.72% F1 score).

(c) Mixed model, trained and tested on both HARP and Omni-MATH (43.88% accuracy, 0.696 mean average error, 44.26% F1 score).

Figure 6: Confusion matrices using the 200 problems per difficulty samples of HARP and Omni-MATH (results on the 10% test split). The first 6 difficulty levels (1 through 6) are included for a total of 1200 problems.

HARP model being unable to predict difficulties above 6, and the Omni-MATH model predicting difficulties above 6 (which are not in HARP).

The confusion matrices for a selection of sub-experiments (Figure 6) shows that, when trained and tested on separate datasets, ModernBERT classifies harder problems as easier, and particularly for the 200 Omni-MATH model on the 200 HARP test set, also classifies easier problems as harder (suggesting that HARP problems are harder to predict). These effects are weaker when ModernBERT is trained on both datasets, resulting in a stronger diagonal in the confusion matrix.

Testing ModernBERT on the Putnam problems shows contrasting results based on which dataset was used during training. When training on Putnam, the model has a bias to over-predict difficulty 9, whereas when trained on Omni-MATH (problems with difficulty in $\{7, 8, 9\}$), the model under-predicts difficulty 9 (Figure 7). See Appendix E for ModernBERT confusion matrices covering the full set of splits we performed over our data.

Overall, we find that ModernBERT has slight drift toward the mean, similar to our initial experimentation with prompting LLMs to do the same task. While LLM prompting typically resulted in a normal distribution of problem difficulty (which is true of Omni-MATH according to Gao et al. (2025)), not all benchmarks follow a normal distribution (e.g. HARP).

By using two different, but related signals (LLM metrics from inference and ModernBERT prediction accuracy), we have shown that HARP and Omni-MATH have quite a few differences. Notably, we show that HARP has some stronger correlations in the LLM metrics than Omni-MATH,

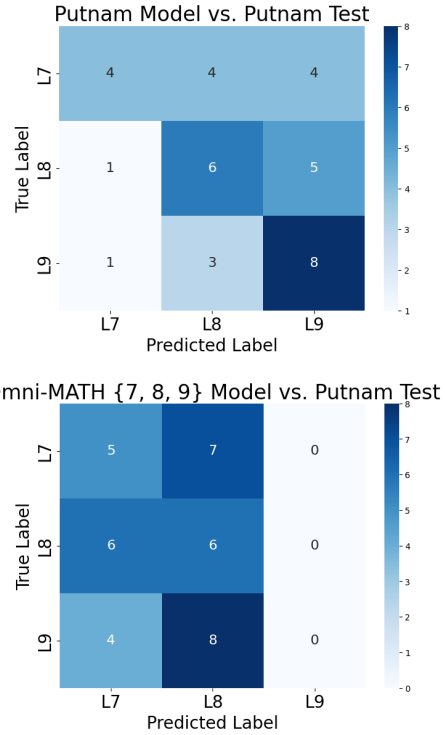


Figure 7: Confusion matrices when evaluating the Putnam and Omni-MATH {7, 8, 9} models on the Putnam test set. Putnam over-predicts difficulty level 9 (50.00% accuracy, 0.639 mean average error, 49.21% F1 score), whereas Omni-MATH under-predicts it (30.56% accuracy, 0.806 mean average error, 24.47% F1 score).

such as the number of tokens in generated solutions, which may indicate that difficulty in HARP is more reflective of "effort" required, i.e. longer solutions. And to contrast, Omni-MATH problems are easier for ModernBERT to predict, reflecting difficulty in Omni-MATH is more semantic/textual, rather than behavioral.

7 Discussion and Future Work

While we highlight issues with present mathematics benchmarks, as they do not properly assess LLM alignment to a single quantitative expert-defined difficulty score, it is beneficial to discuss future directions based on our findings.

Multi-dimensional Difficulty. A key direction for future work is the development of multi-dimensional difficulty labels that build upon the current difficulty score used in modern benchmarks. As discussed in our work, difficulty is inherently multifaceted, encompassing dimensions such as effort, required conceptual knowledge, and uncertainty during problem solving. A single expert-annotated difficulty level cannot capture these nuanced distinctions. Future benchmarks should look into annotating problems along multiple axes, including procedural effort (number of reasoning steps), uncertainty (similarity to previous knowledge, how many reasoning paths exist), conceptual depth (reliant on a student's current educational level), etc. Such a framework would allow for a precise way to measure alignment between human notions of difficulty and LLM-derived metrics. By explicitly modeling these dimensions, future work could help us find which aspects of difficulty LLMs "understand", and which go unmodeled. While we discovered these potential dimensions of difficulty through discussions, it would be beneficial to look toward education literature to help bridge the gap with present NLP research.

Human vs. LLM Reasoning. Another direction for future research is the systematic study of reasoning steps in both human solutions and LLM-generated solutions. While prior work and our current findings have shown that longer solutions often correlate with higher difficulty, the number of reasoning steps alone may not be a reliable indicator. For instance, a more advanced student (e.g., a student who is currently taking Calc II) may solve a derivative problem in fewer, and more abstract, steps, when compared to a less experienced student (e.g., a student who is beginning Calc I and only knows the limit definition of the derivative). This is more explicit, for example, when asking elementary questions to a reasoning model, for example "What is $9 + 10$?", in which a reasoning model will likely give more than one reasoning step to verify its solution. This suggests that step count is not just a function of problem difficulty, but also of the student (or LLM's) background experience with solving problems of similar nature. Future work could look into richer representations of reasoning beyond a single quantitative measure.

Variations in Reasoning. This naturally motivates the need to study reasoning paths across different populations, rather than one solution by one person (or model) at a time. Rather than treating the solution as having a static baseline, future work should account for differences in how different people approach the same problem, highlighting the different abilities and disabilities that interact with the problem-solving process. Comparing these human reasoning paths to those generated by LLMs may reveal whether models more closely align to a student, or an educator, or perhaps both depending on the context and prompting used. Such analyses could provide deeper insight into how LLMs internalize mathematical reasoning when compared to humans, which would help inform future designs.

8 Conclusion

Using LLM inferences from the HARP (Yue et al., 2024) and Omni-MATH (Gao et al., 2025) benchmarks, we explored several signals from LLMs to determine their potential relation to human notions of difficulty (in the form of difficulty level labeled by experts). Our findings suggest that solution accuracy may not be the only LLM metric that aligns to human-annotated difficulty of math problems, as log probabilities and solution lengths also correlate moderately (and sometimes more strongly) to difficulty level. We also support the findings from Luo et al. (2025) that geometric spatial problems are more difficult for LLMs to accurately answer. Additionally, we explored the relation between the length of a problem and the length of its LLM solution, finding very weak correlation.

Building on the Omni-MATH dataset, which contains problems from the Putnam competition, we collected historical human scores from each competition, enabling us to investigate the alignment between LLM metrics, human accuracy, and expert-annotated difficulty scores, finding strong correlation between the latter two and weak correlations with LLM metrics.

Finally, we explored token and sentence-level semantics by training several instances of ModernBERT across HARP, Omni-MATH, and Putnam problems. When conducting cross evaluations between HARP and Omni-MATH, we find that model performance always drops, suggesting com-

petitions within the two benchmarks may have different relations to expert-labeled difficulty scores. While training ModernBERT on both benchmarks has higher accuracy than our cross evaluations, it does not recover the full accuracy of any model trained solely on one dataset.

We hope our work leads to further inquiry into the current nuances of LLM alignment to difficulty for mathematics benchmarks, as well as a more fine-grained approach that represents the wide-range of human notions of math difficulty (rather than a single scale).

Limitations

While we can explore LLM alignment to a single measure of human annotated difficulty, we find that we are unable to truly explore LLM alignment to all human intuitions of difficulty, which likely has many facets (such as the structure of the problem, or the list of skills necessary to solve). We are not altering any problems to control for problem or solution length, and therefore we cannot determine any causal link between human-annotated difficulty, problem length, or solution length. Our results are limited only to the 6 LLMs selected, which do not include API models from multiple families (e.g. Claude or Meta), or reasoning models (e.g. OpenAI's o3 model).

Acknowledgments

We would like to thank our many anonymous reviewers for their helpful feedback, as well as UM-Flint College of Innovation & Technology for funding the Summer Undergraduate Research Experience (SURE) program in summer 2025 that supported this work. We give special thanks to Anne Jonas and Maeve McLaughlin for the many early and thought provoking discussions that led to this work.

References

Art of Problem Solving. 2025. Aops wiki: Competition ratings. https://artofproblemsolving.com/wiki/index.php/AoPS_Wiki:Competition_ratings. Accessed: 2025-08-21.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024a. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyi Duan, Hengnian Gu, Yuan Ke, and Dongdai Zhou. 2024b. Ebert: A lightweight expression-enhanced large-scale pre-trained language model for mathematics education. *Knowledge-Based Systems*, 300:112118.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Anthony W.F. Lao and Philip I.S. Lei. 2023. Subject classification and difficulty ranking of math problems. In *2023 IEEE 11th International Conference on Information, Communication and Networks (ICICN)*, pages 844–850.

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. MathFish: Evaluating language model math reasoning via grounding in educational curricula. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5644–5673, Miami, Florida, USA. Association for Computational Linguistics.

Shixian Luo, Zezhou Zhu, Yu Yuan, Yuncheng Yang, Lianlei Shan, and Yong Wu. 2025. Geogrambench: Benchmarking the geometric program reasoning in modern llms. *Preprint*, arXiv:2505.17653.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2025-08-21.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. *Preprint*, arXiv:2410.02707.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *Preprint*, arXiv:2105.00377.

Piotr Piękos, Mateusz Malinowski, and Henryk Michalewski. 2021. Measuring and improving BERT's mathematical abilities by predicting the order of reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 383–394, Online. Association for Computational Linguistics.

Benjamin Plaut, Nguyen X. Khanh, and Tu Trinh. 2025. Probabilities of chat llms are miscalibrated but still predict correctness on multiple-choice q&a. *Preprint*, arXiv:2402.13213.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2023. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *Preprint*, arXiv:2106.07340.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

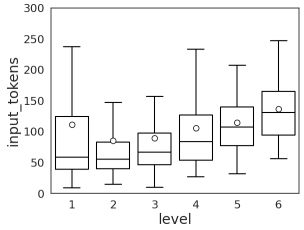
Albert S. Yue, Lovish Madaan, Ted Moskovitz, DJ Strouse, and Aaditya K. Singh. 2024. Harp: A challenging human-annotated math reasoning benchmark. *Preprint*, arXiv:2412.08819.

Appendix A Model-related Statistics

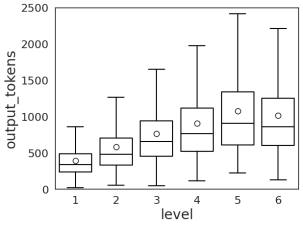
For brevity, we expand on Figure 2 by including the rest of the LLM metric distributions grouped by difficulty level for HARP and Omni-MATH problems (Figures 8 and 9).

Appendix B LLM Metrics Correlation Heatmaps

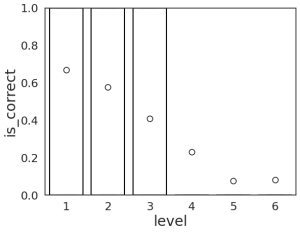
For brevity, we expand on Figures 3 and 4 by showing the separate Pearson correlation heatmaps of the LLM metrics for each model. Across models, we find similar patterns for `is_correct`, `level`, `min_logprob`, `input_tokens`, and `output_tokens` (Figures 10 and 11). Note that `Mathstral-7B-v0.1` does not use `\boxed`; when aggregating `min_logprob_boxed`, this LLM is ignored.



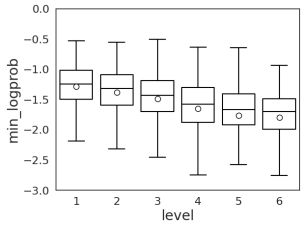
(a) Distribution of `input_tokens`.



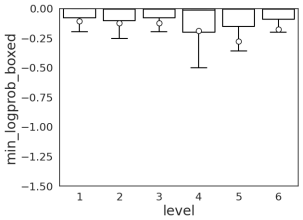
(b) Distribution of `output_tokens`.



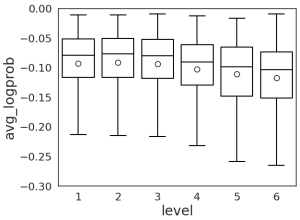
(c) Distribution of `is_correct`.



(d) Distribution of `min_logprob`.

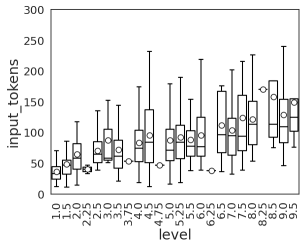


(e) Distribution of `min_logprob_boxed`.

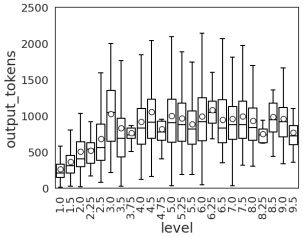


(f) Distribution of `avg_logprob`.

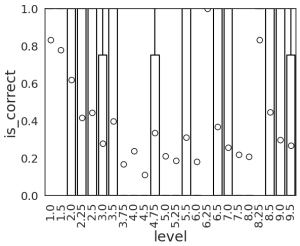
Figure 8: Distributions of model-related statistics grouped by difficulty level for HARP problems.



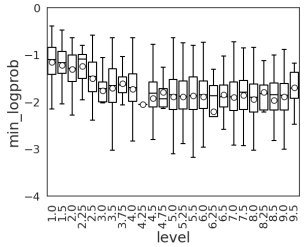
(a) Distribution of `input_tokens`.



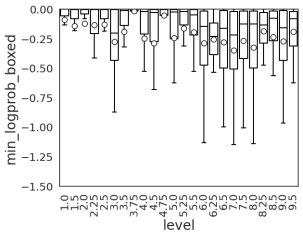
(b) Distribution of `output_tokens`.



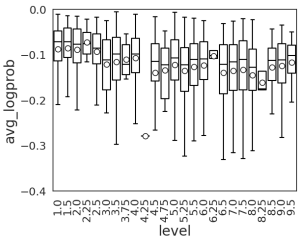
(c) Distribution of `is_correct`.



(d) Distribution of `min_logprob`.

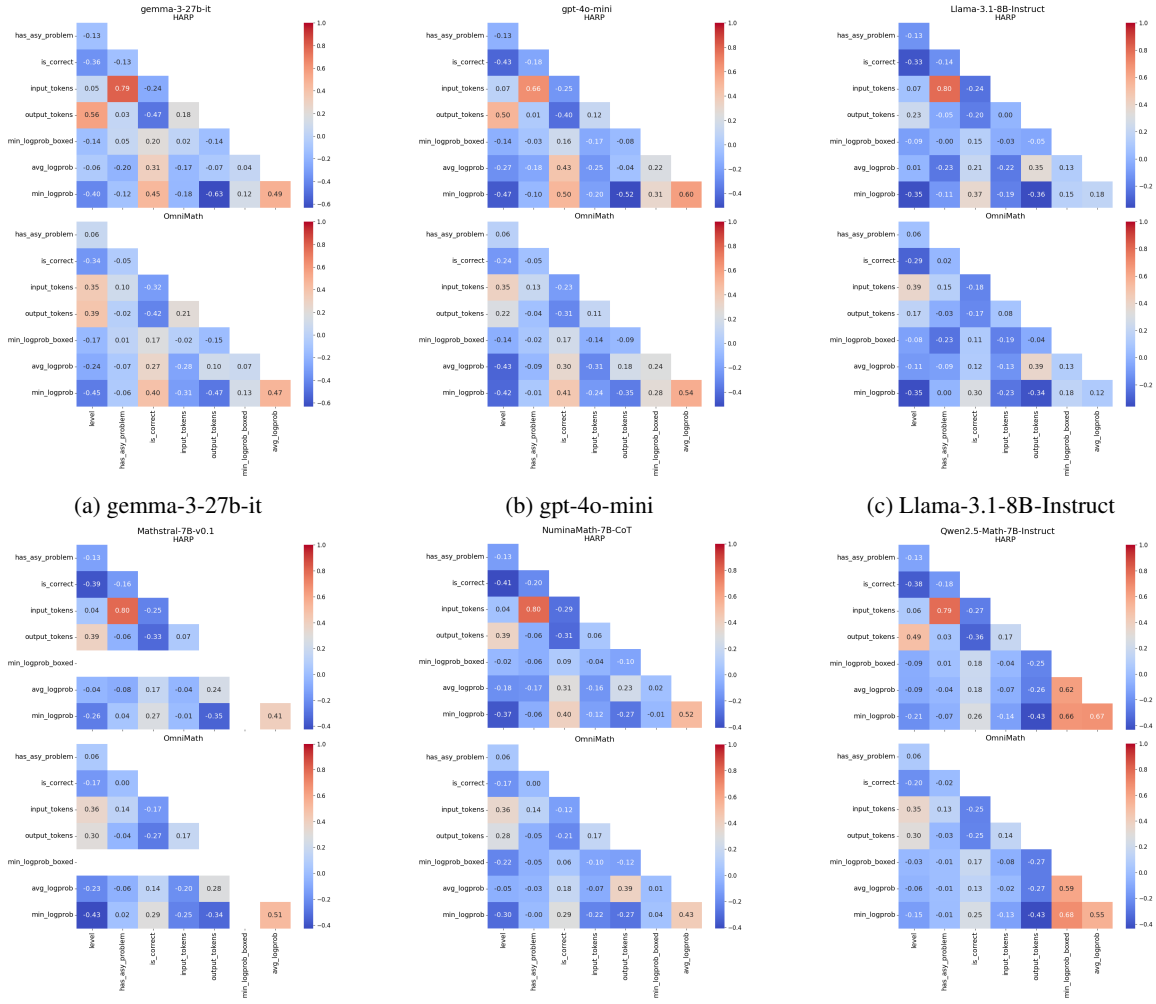


(e) Distribution of `min_logprob_boxed`.



(f) Distribution of `avg_logprob`.

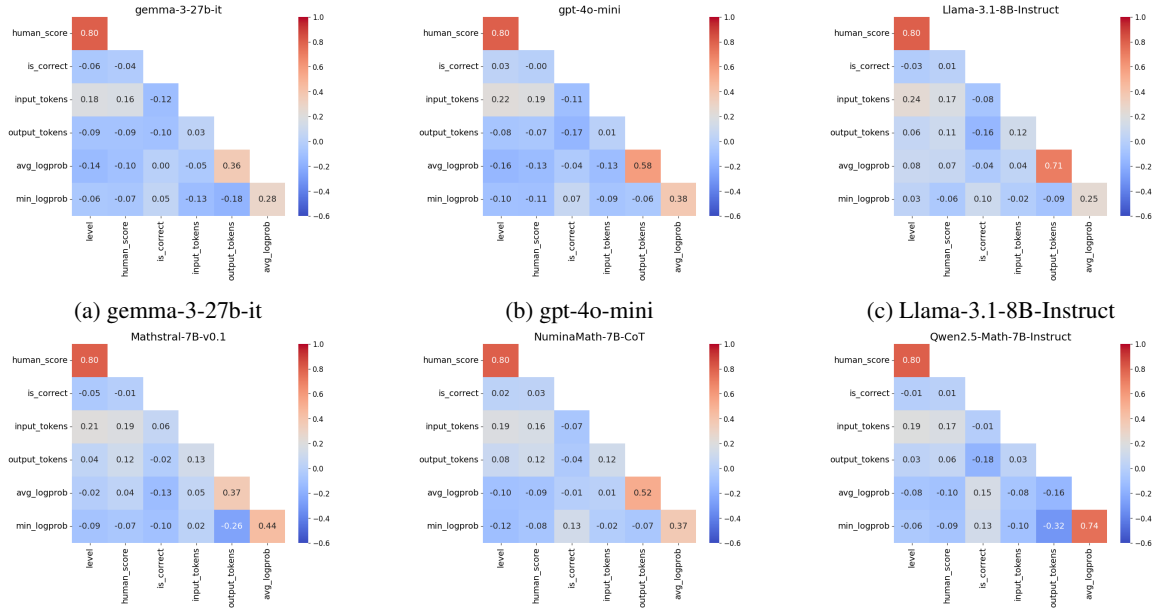
Figure 9: Distributions of model-related statistics grouped by difficulty level for Omni-MATH problems.



(a) gemma-3-27b-it (b) gpt-4o-mini (c) Llama-3.1-8B-Instruct

(d) Mathstral-7B-v0.1 (e) NuminaMath-7B-CoT (f) Qwen2.5-Math-7B-Instruct

Figure 10: Pearson correlation heatmap of the variables from the HARP and Omni-MATH problems, 6 models.



(a) gemma-3-27b-it (b) gpt-4o-mini (c) Llama-3.1-8B-Instruct

(d) Mathstral-7B-v0.1 (e) NuminaMath-7B-CoT (f) Qwen2.5-Math-7B-Instruct

Figure 11: Pearson correlation heatmap of the variables from the Putnam problems, 6 models.

Appendix C OLS Regression on LLM Metrics

We conduct a series of OLS regression tests on difficulty level using our LLM metrics to control for different metrics and see how they interact. Overall, we find that adjusted R^2 is low, especially for Omni-MATH, and find that `input_tokens` and `min_logprob` has stronger correlation for difficulty level in Omni-MATH than in HARP (Table 2). Future work could involve constructing a dataset where problem length and solution length is held constant, which could then be used in our ModernBERT experiments (Section 5) to control for interactions between `min_logprob` and `ouput_tokens`.

Linear Model	**Adjusted R^2**	
	HARP	**Omni-MATH**
L ∼ avg	0.007	0.015
L ∼ min	0.075	0.061
L ∼ output	0.175	0.064
L ∼ output + min	0.197	0.103
L ∼ asy	0.016	0.003
L ∼ asy + output + min	0.216	0.106
L ∼ input	0.103	0.103
L ∼ input + output + min	0.197	0.176
L ∼ corr	0.135	0.047
L ∼ corr + asy	0.172	0.050
L ∼ corr + min	0.161	0.085
L ∼ corr + output	0.252	0.093
L ∼ corr + output + min	0.257	0.118
L ∼ corr + output + min + asy	0.289	0.121
L ∼ all but avg	0.302	0.184
L ∼ all	0.302	0.184

Table 2: Comparison of adjusted R^2 when predicting `level` (L) with `avg_logprob` (avg), `min_logprob` (min), `output_tokens` (output), `has_asy_problem` (asy), `input_tokens` (input), and `is_correct` (corr).

Appendix D Humans vs. LLMs on Putnam Competition

For brevity, we expand on Figure 5 by including the individual LLM accuracies at each difficulty level for problems in the Putnam competition (Figure 12). We find that each LLM follows the pattern of constant accuracy as difficulty level increases.

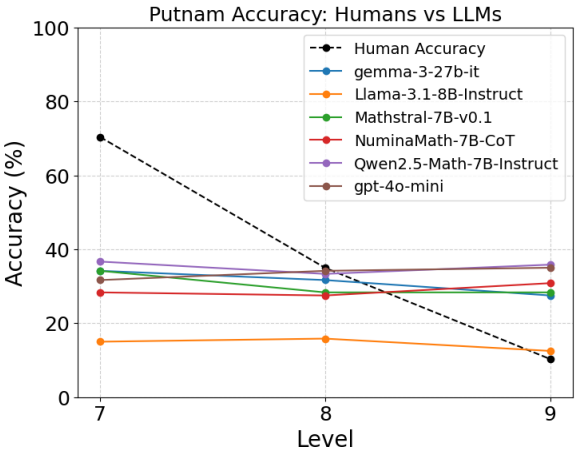


Figure 12: Accuracy for humans and LLMs when solving Putnam problems.

Appendix E ModernBERT Confusion Matrices

For brevity, we expand on Figure 6 by including ModernBERT confusion matrices covering the full set of splits we performed over our data (Figure 13). On the left, we have the models trained using the 200 problems per difficulty level, tested on their own test set (instead of cross-tested as in Figures 6a and 6b). On the right, we have the models trained using the full dataset (instead of the balanced 200 problems per difficulty level), tested on their own test set. In the middle, we have the models trained on the full dataset, tested on their opposite model dataset. The top row contains models trained on the HARP dataset, whereas the bottom row contains models trained on the Omni-MATH dataset.

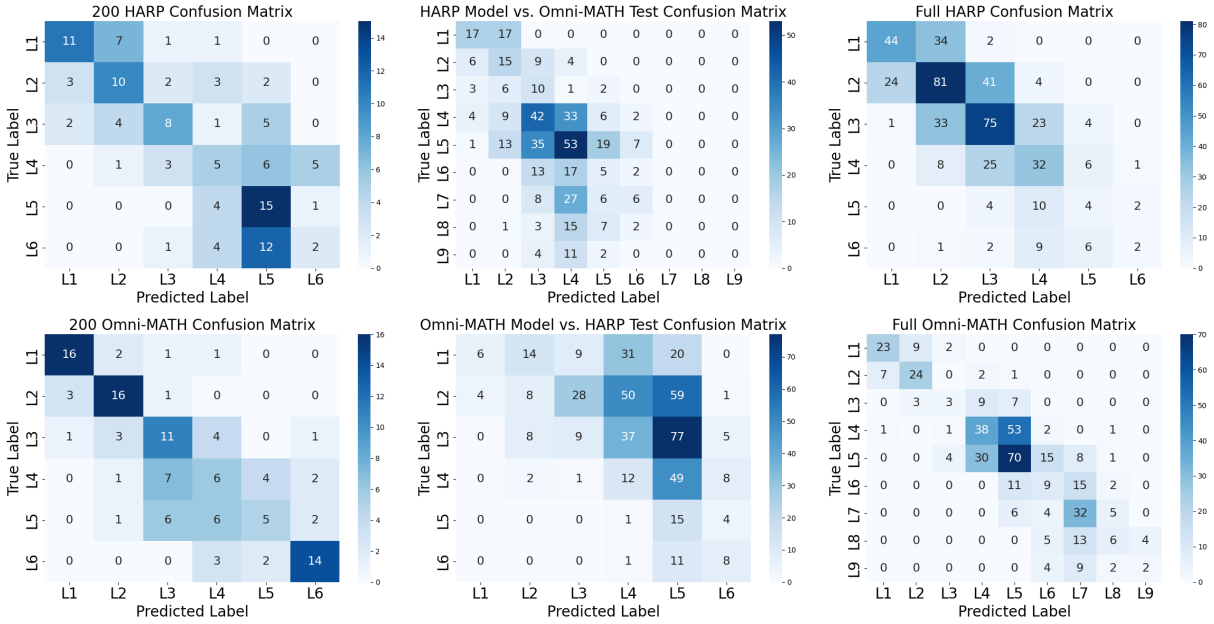


Figure 13: ModernBERT confusion matrices