Putnam-MATH: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning

Aryan Gulanti
Stanford University
aryangul@stanford.edu

Eric Chen
Stanford University
brando9@stanford.edu

Brando Miranda
Stanford University
brando9@stanford.edu

Kai Fronsdal
Stanford University
kaif@stanford.edu

Emily Xia
Stanford University
emxia@stanford.edu

Bruno Dumont
Stanford University
bdumont@stanford.edu

May 15, 2024

Abstract

The advent of Large Language Models (LLMs) has led to an unprecedented speed of improvement in AI capabilities. For instance, within three years of introducing the MATH dataset for mathematical reasoning in 2021 , models have attained an accuracy of 85% in 2024 – a 12-fold improvement from the original 6.9%. In addition, this achievement is merely 10% short of the 95% accuracy level reported for a three-time International Mathematical Olympiad (IMO) gold medalist. Therefore, as AI models become more capable and quickly begin to approach ceiling performances on established benchmarks, there arises a need for more **challenging** and **long-lasting** evaluation benchmarks. Therefore, we introduce the Putnam-MATH dataset, a unique collection of higher level mathematics problems that require expert-level understanding solving. Our dataset is characterized by its challenging nature, with participants – aspiring professional mathematicians (undergraduate participants) – scoring a median of zero in the Putnam competition in 2008. In addition, numerous Putnam Fellows have achieved prominence in mathematics and other disciplines, including four who have won the Fields Medal — Terence Tao, John Milnor, David Mumford, and Daniel Quillen — and two who have received Nobel Prizes in Physics, Richard Feynman and Kenneth Wilson. Motivated by the risks of data contamination and goal to make a long-lasting benchmark in the era of fast pace of AI progress, we introduced a functional variation to our dataset. This variation modifies variable names and constants in the problems, keeping the conceptual and reasoning aspects intact, which helps in avoiding memorization by models with the potential *infinite* variations for some problems. Our evaluations demonstrate our benchmark is indeed difficult: GPT-4 gets 14/192 questions correctly, a specialized mathematics model like DeepSeekMath-7B gets 8/192, and popular 7B open-source models like LLama3-8B score 6/192. For our dataset's functional variation, the numbers are even more stringent, with GPT-4 scoring 2/35, DeepSeekMath-7B 3/35, and LLama3-8B 0/35, further validating the challenging nature of our tests. The challenging nature of our datasets, both in their original and varied forms, not only tests the limits of current AI capabilities but also paves the way for new research directions in AI to push the boundaries of deep mathematical reasoning.

1 Methods

1.1 Putnam Static Dataset

To create a challenging dataset for testing the mathematical problem-solving capabilities of large language models (LLMs), a set of 192 questions was curated from the Putnam Mathematical Competition problems posed between 1995 and 2023. Drawing inspiration from the MATH dataset by [Hendrycks et al.(2020)Hendrycks, Burns, Basa we use latex boxing to capture the answer string for evaluating mathematical understanding in LLMs, e.g., `boxed{2n}`. Boxed answers are crucial for facilitating automated evaluation and ensuring that LLMs generate precise, unambiguous responses but have the drawback of not directly actual reasoning string. However, due to the intricate nature of the problems and their solutions, not all Putnam questions could be directly

Table 1: Performance of various models on the original static and automatically generated variations of the Putnam-MATH benchmark.

Model	Original Static		Automatically Generated	
	Score	**Percentage**	**Score**	**Percentage**
GPT-4o	21/192	10.9%	3/35	8.57%
GPT-4-turbo	14/192	7.29%	2/35	5.71%
GPT-3.5-turbo	6/192	3.13%	0/35	0%
DeepSeekMath7bInstruct	8/192	4.17%	3/35	8.57%
Mistral7B	6/192	3.13%	0/35	0%
LLama3-7B	6/192	3.13%	0/35	0%
Gemma2B	6/192	3.13%	0/35	0%

converted into a format suitable for testing LLMs, as some answers were provided in a prose format rather than a concise, and boxable answer.

1.2 Putnam Variation Dataset

As previous years' Putnam problems are available on the Putnam website, there is a possibility that models have been fed those problems as training data and would thereby have an artificially high accuracy. In order to properly test the mathematical ability of LLMs and in order to compile a larger testing dataset, we include functional variations on these problems.

Our functional variations cover a wide range of changes which are detailed below.

1. We significantly alter a problem's structure e.g. changing "Find solutions to $f(x) = \sin(x)$" to "Find solutions to $f(x) = \cos(x)$".
2. We adjust constants within the problem (e.g. changing "Evaluate $\sum_{n=1}^{2022} a_n$" to "Evaluate $\sum_{n=1}^{5000} a_n$ ").
3. We change variable names (e.g. changing "Solve $x + 5 = 7$" to "Solve $b + 5 = 7$").

We currently have functional variations of seven different Putnam questions. Each variation is capable of generating an infinite number of unique, equal-difficulty questions. For our model evaluation, we generated five unique questions per variation, giving us a total of 35 variation questions.

2 Results

References

[Hendrycks et al.(2020)Hendrycks, Burns, Basart, Zou, Mazeika, Song, and Steinhardt] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

3 Supplementary Material

4 Conclusion

In conclusion, the Putnam-MATH dataset and its functional variation provide a rigorous and long-lasting benchmark for evaluating the mathematical reasoning capabilities of AI models. By introducing a challenging and diverse set of problems, we aim to drive the development of more advanced AI systems that can excel in complex mathematical problem-solving. Our work serves as a foundation for future research in pushing the boundaries of AI capabilities in mathematics and beyond.

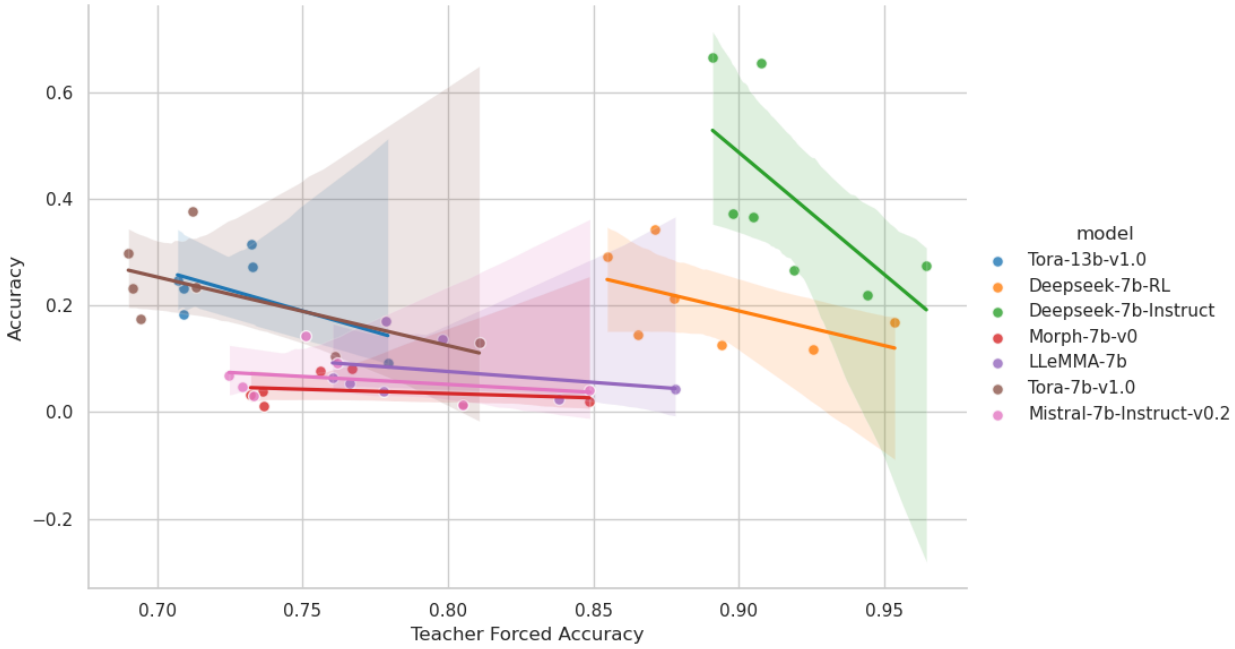

Figure 1: **Demonstrates the moderate correlation between boxed answers accuracy and teacher forced accuracy on the MATH math data set.** Shaded areas correspond to 95% confidence intervals. We evaluated the four model across seven mathematical topics of the MATH data set: Number Theory, Intermediate Algebra, Algebra, Geometry, Precalculus, Counting and Probability, Prealgebra. Using teacher forcing results a clear negative correlation conditioned on model, but over all is not reliable.

4.1 Calculating Aggregated ROSCOE

ROSCOE is a collection of metrics each designed to measure a different aspect of reasoning. In the original paper, the authors gave no way of combining the different metrics into an aggregate score of correctness. The focus of our work is to benchmark model performance which requires a single comparable metric for each dataset. In total, there are 19 different base metrics in ROSCOE, which we label m_i for $1 \leq i \leq 19$; computing each of these on a large dataset is time consuming, so ideally we could restrict ourselves to a few of the most useful metrics. We learn a simple linear combination of the metrics and employ L_1 regularization to promote sparsity. Thus we are trying to find the coefficients α_i to construct our aggregated metric M as follows

$$M = \sum_{i=1}^{k} \alpha_i m_i$$

where α_i are sparse. We find that learning α_i on all datasets results in a correlation with boxed accuracy of 0.919.

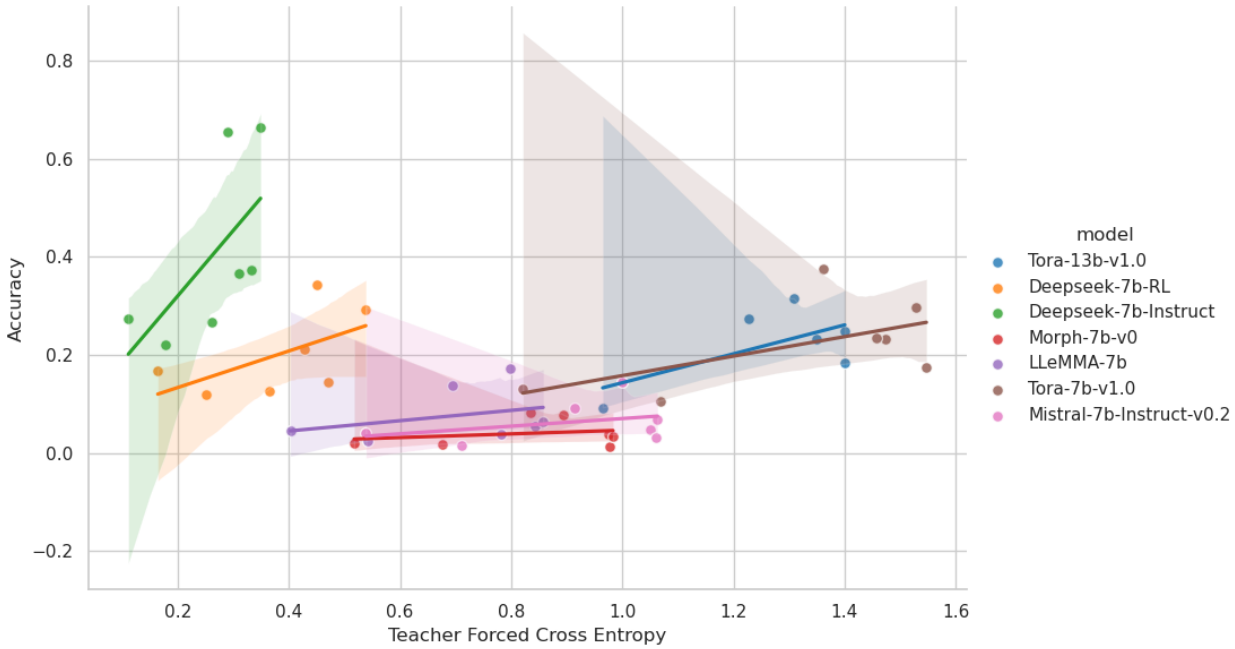

Figure 2: **Demonstrates the moderate correlation between boxed answers accuracy and teacher forced cross entropy on the MATH math data set.** Shaded areas correspond to 95% confidence intervals. We evaluated the four model across seven mathematical topics of the MATH data set: Number Theory, Intermediate Algebra, Algebra, Geometry, Precalculus, Counting and Probability, Prealgebra. Using teacher forced cross entropy results in more comparable results across datasets and models, with trends between models and datasets matching, but still too dependent on confounding factors.

model	TFA	TFCE
Deepseek-7b-Instruct	0.439272	0.399167
Deepseek-7b-RL	0.285773	0.317733
LLeMMA-7b	0.100595	0.104666
Mistral-7b-Instruct-v0.2	0.097707	0.131686
Morph-7b-v0	0.060026	0.053734
Tora-13b-v1.0	0.311476	0.389184
Tora-7b-v1.0	0.376615	0.331863

Figure 3: R^2 **between metric and ground truth accuracy for different models.** Results are fairly similar between TFA and TFCE.

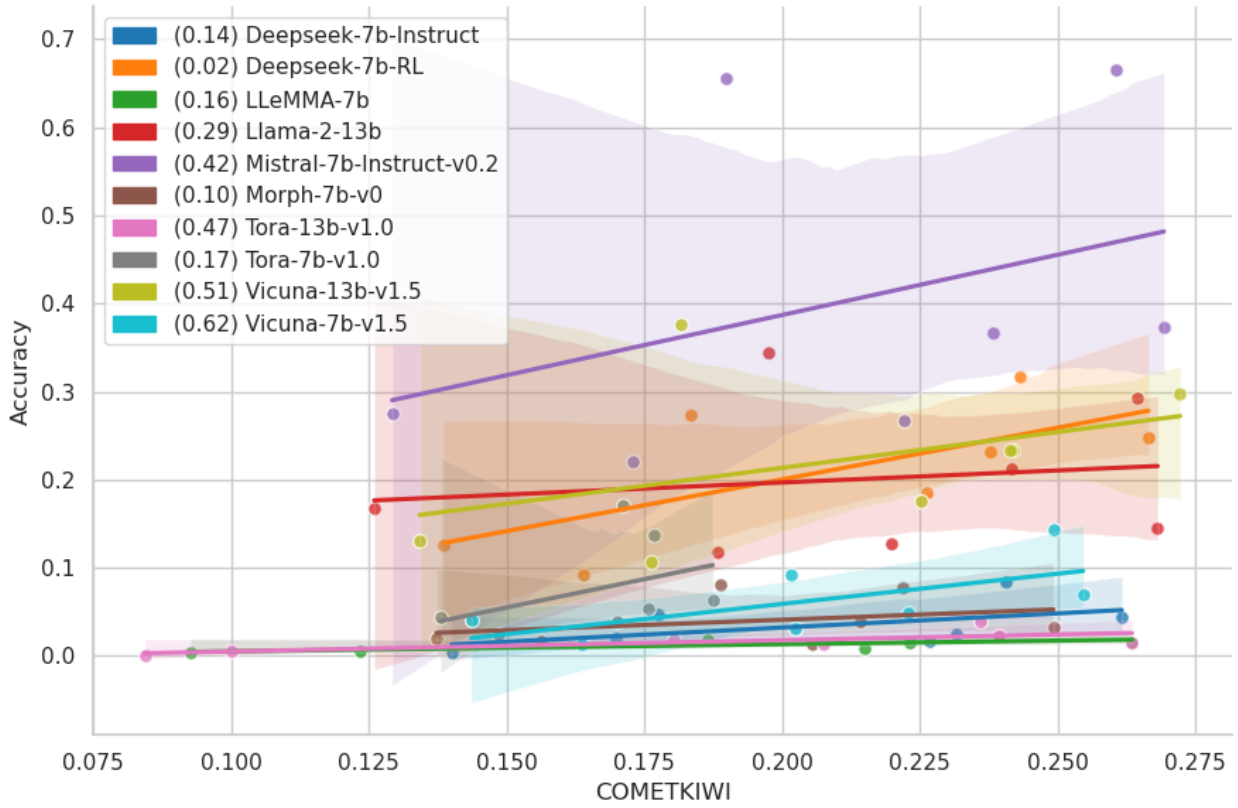

Figure 4: **Demonstrates the low to medium correlation between boxed answers accuracy and COMETKIWI on the MATH math data set.** Shaded areas correspond to 95% confidence intervals. R^2 is given in parentheses in the legend. We evaluated the multiple model across seven mathematical topics of the MATH data set: Number Theory, Intermediate Algebra, Algebra, Geometry, Precalculus, Counting and Probability, Prealgebra. Using COMETKIWI results is strong correlations conditioned on model, but results between models are still not comparable. (Note: COMETKIWI does not use the reference answer and is fine-tuned for language translation, not mathematical accuracy.)

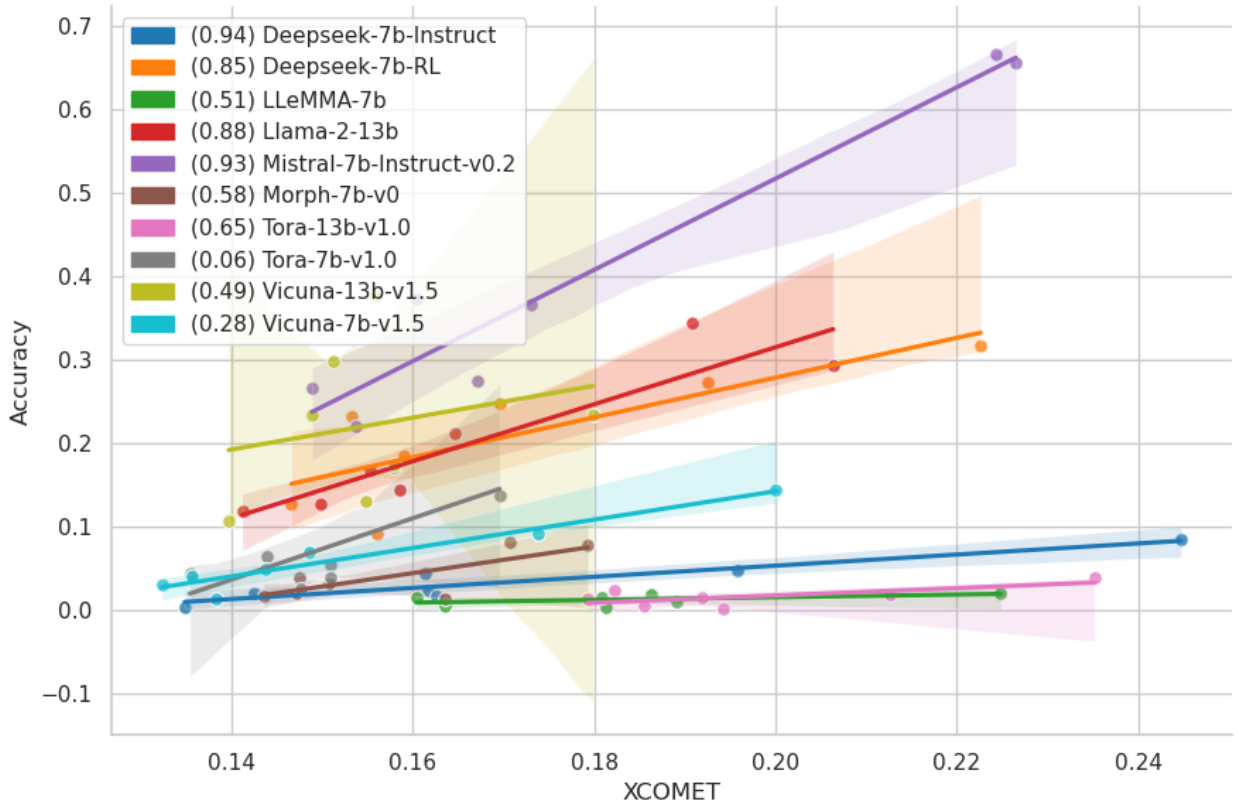

Figure 5: **Demonstrates the medium to strong correlation between boxed answers accuracy and XCOMET on the MATH math data set.** Shaded areas correspond to 95% confidence intervals. R^2 is given in parentheses in the legend. We evaluated the multiple model across seven mathematical topics of the MATH data set: Number Theory, Intermediate Algebra, Algebra, Geometry, Precalculus, Counting and Probability, Prealgebra. Using XCOMET results is strong correlations conditioned on model, but results between models are still not comparable. (Note: XCOMET uses the reference answer and is fine-tuned for language translation, not mathematical accuracy.)

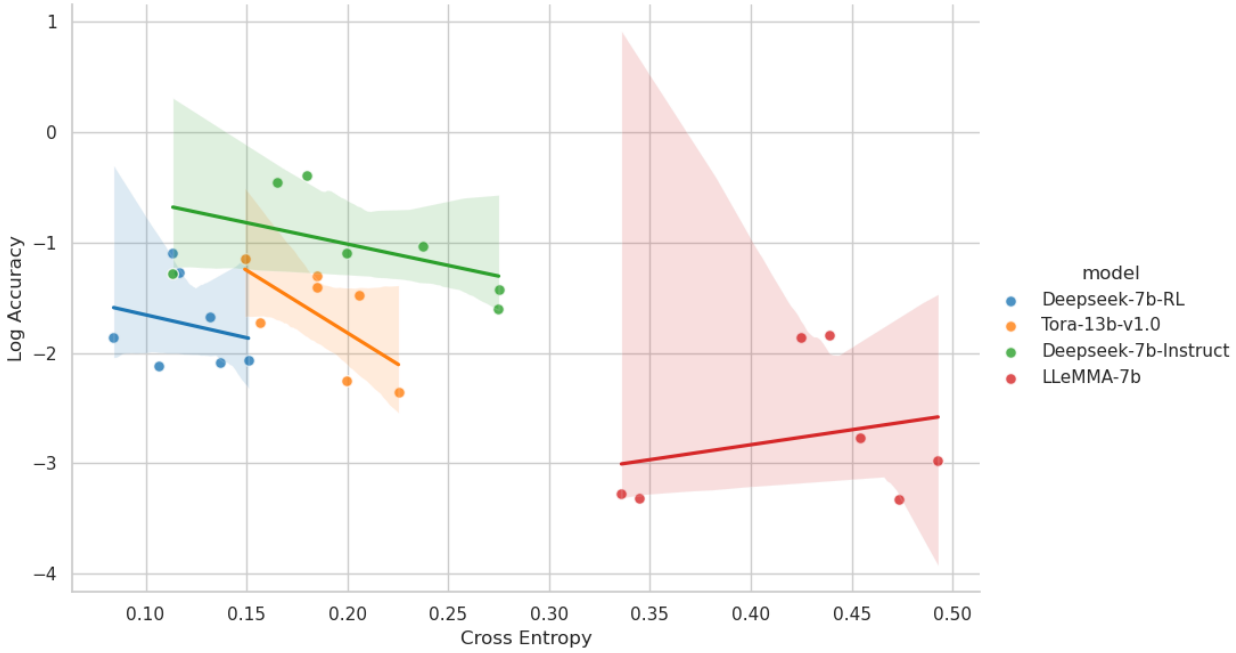

Figure 6: **Demonstrates the low correlation of accuracy using the boxed answers metric vs autoregressive cross entropy, both on the MATH math data set.** EleutherAI-llema $R^2 = 0.0631$ (red line), Deepskee-ai-deepseek-math-7b-instruct (green line) $R^2 = 0.248$, Deepseek-ai-deepskee-mth-7b-rl (blue line) $R^2 = 0.0491$, LLM-agents-tora-13b-v1.0 (orange) $R^2 = 0.422$. Shaded areas correspond to 95% confidence intervals. We evaluated the four model across seven mathematical topics of the MATH data set: Number Theory, Intermediate Algebra, Algebra, Geometry, Precalculus, Counting and Probability, Prealgebra. We hypothesize this low correlation suggests perplexity as a bad surrogate for boxed answer accuracy for comparison between different models (preferred metric, though harsh) used in the MATH data set math and therefore display the low).

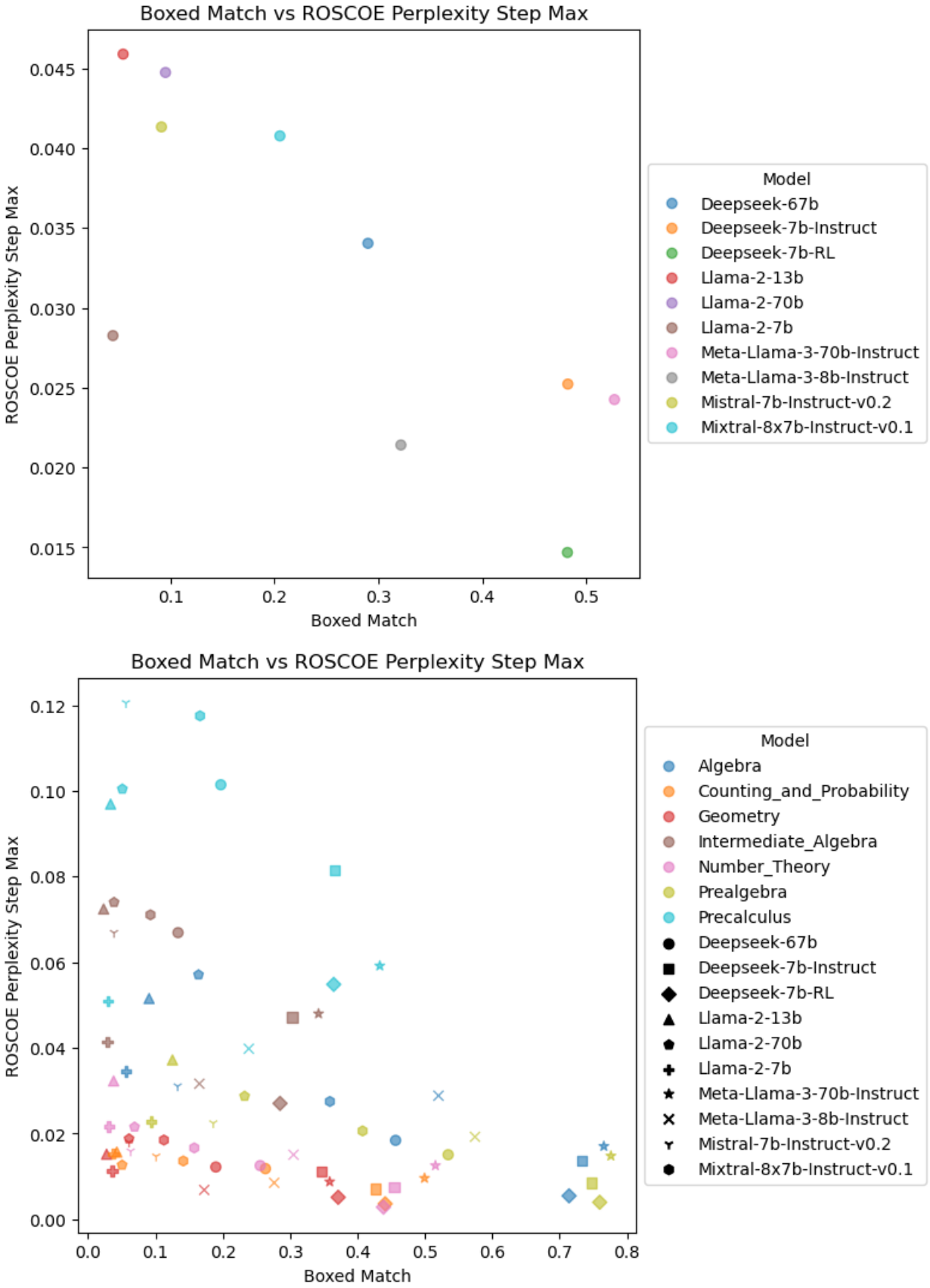

Figure 7: Relationship between ROSCOE Perplexity Step Max and Boxed Accuracy. (top) metrics are averaged over all datasets; (bottom) metrics are computed per dataset.

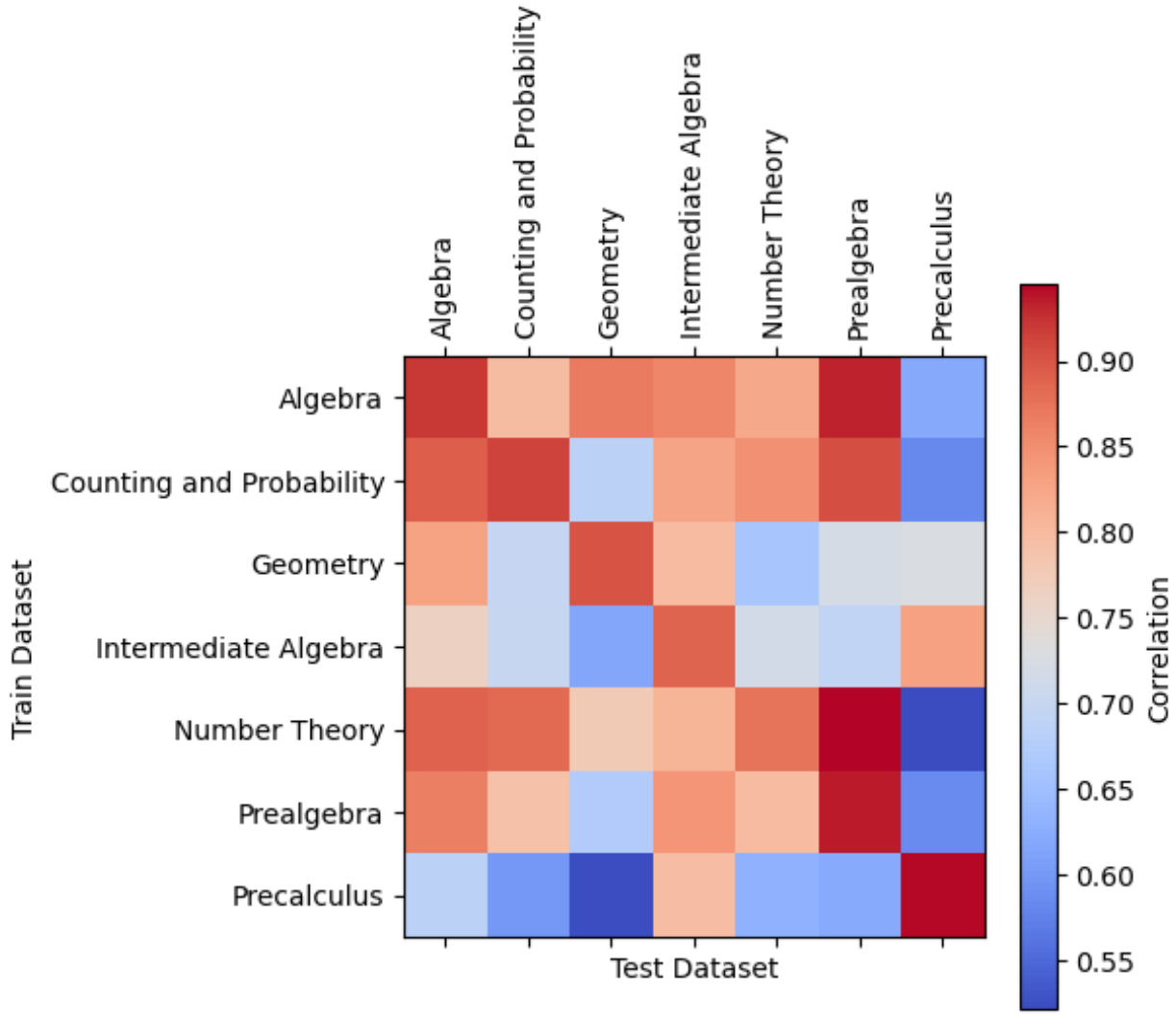

Figure 8: Generalization performance of Aggregated ROSCOE fit on a single dataset and the remaining datasets. Each row corresponds to the dataset Aggregated ROSCOE was fit on and each column corresponds to the dataset correlation to boxed accuracy was calculated with.

ROSCOE is a collection of metrics each designed to measure a different aspect of reasoning. In the original paper, the authors gave no way of combining the different metrics into an aggregate score of correctness. The focus of our work is to benchmark model performance which requires a single comparable metric for each dataset. In total, there are 19 different base metrics in ROSCOE, which we label m_i for $1 \leq i \leq 19$; computing each of these on a large dataset is time consuming, so ideally we could restrict ourselves to a few of the most useful metrics. We learn a simple linear combination of the metrics and employ L_1 regularization to promote sparsity. Thus we are trying to find the coefficients α_i to construct our aggregated metric M as follows

$$M = \sum_{i=1}^{k} \alpha_i m_i$$

where α_i are sparse. We find that learning α_i on all datasets results in a correlation with boxed accuracy of 0.919.

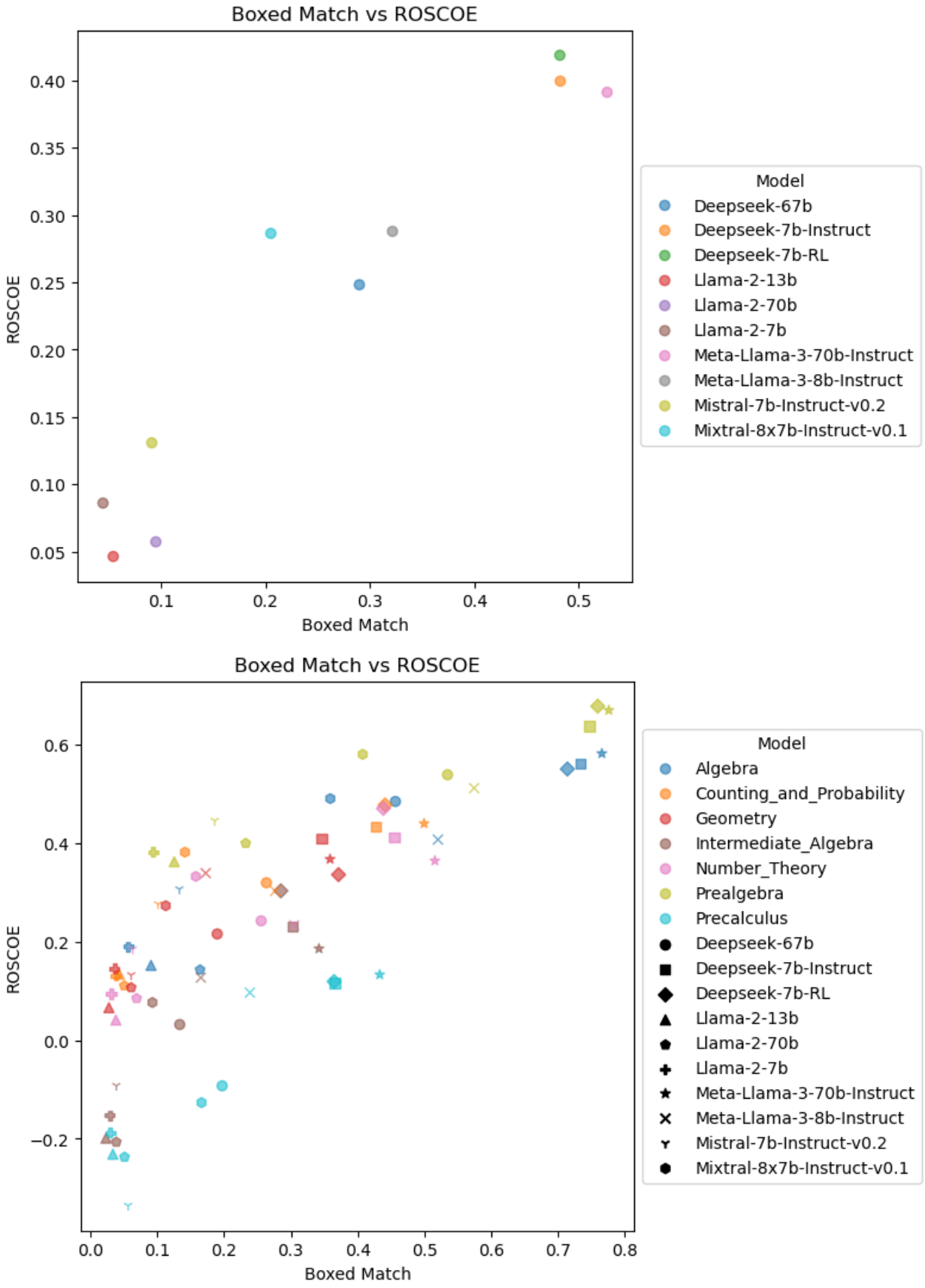

Figure 9: Relationship between Aggregated ROSCOE and Boxed Accuracy. Aggregated ROSCOE refers to the linear model of the 5 most important ROSCOE metrics. (top) metrics are averaged over all datasets; (bottom) metrics are computed per dataset.

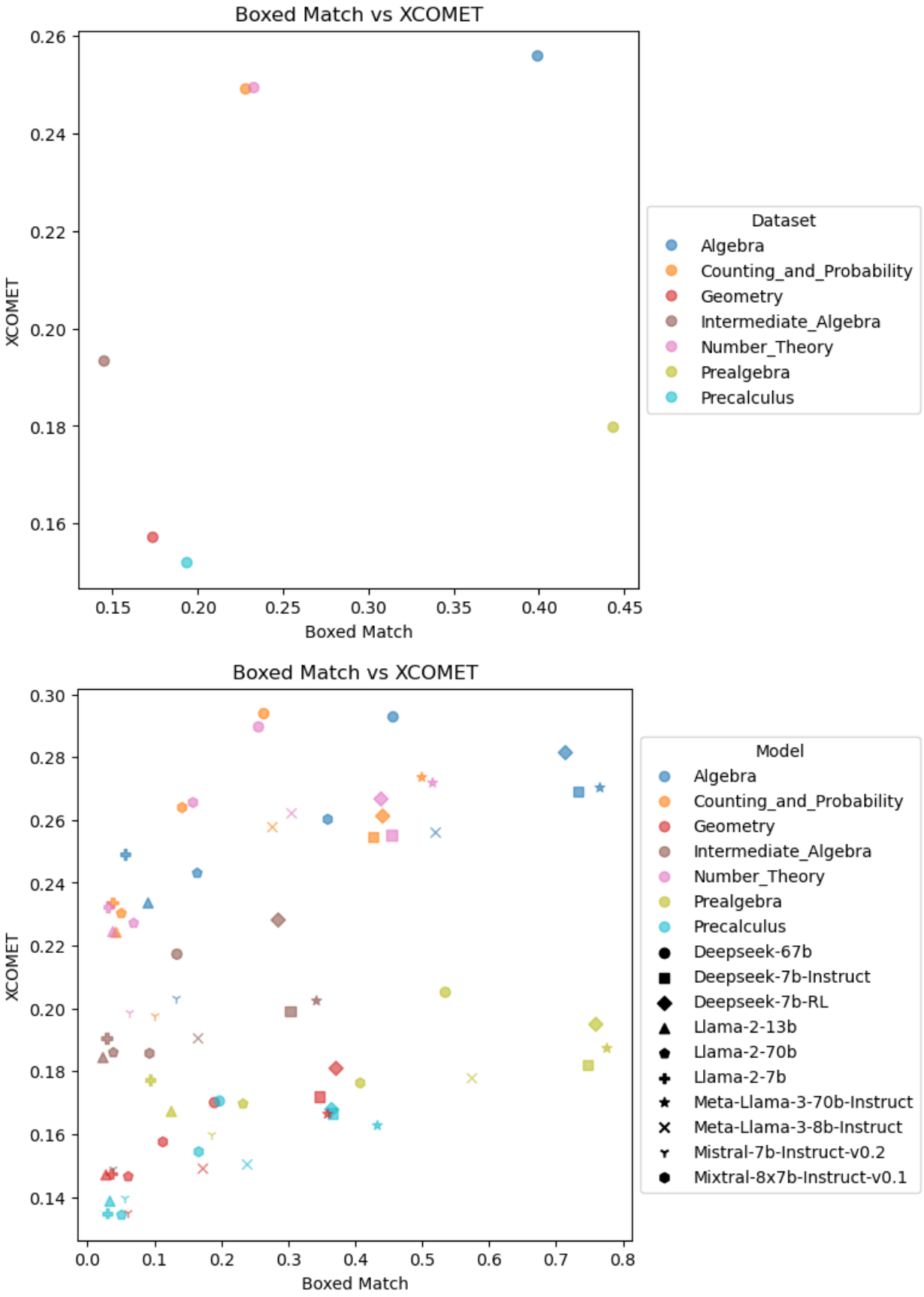

Figure 10: Relationship between XCOMET and Boxed Accuracy. (top) metrics are averaged over all datasets; (bottom) metrics are computed per dataset.