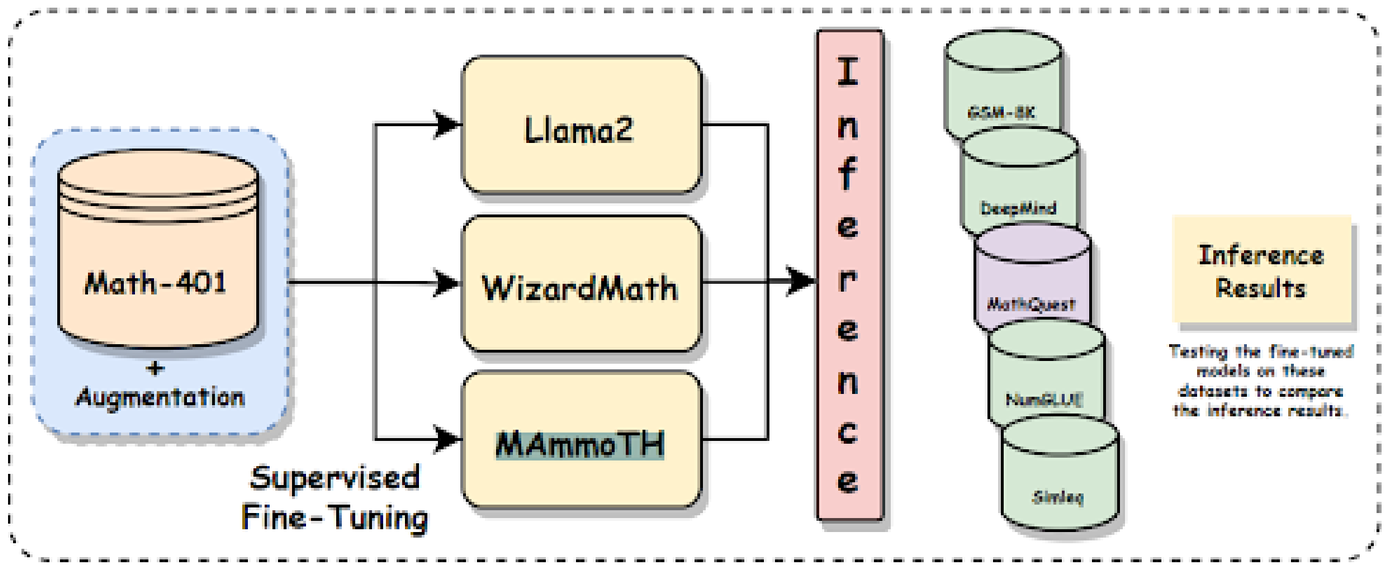
Mathify: Evaluating Large Language Models on Mathematical Problem Solving Tasks

Avinash Anand, Mohit Gupta, Kritarth Prasad, Navya Singla, Sanjana Sanjeev,
Jatin Kumar, Adarsh Raj Shivam, Rajiv Ratn Shah
Indraprastha Institute of Information Technology, Delhi

ABSTRACT

This paper endeavors to tackle the challenges posed by mathematical problem-solving within the context of LLMs. To this end,

- **We introduce MathQuest, a comprehensive mathematics dataset meticulously curated from the 11th and 12th standard Mathematics NCERT textbooks. This dataset spans various levels of mathematical complexity and encompasses a wide array of mathematical concepts.**
- **To equip Large Language Models (LLMs) with the ability to solve these intricate problems, we conduct fine-tuning on this dataset.**
- **We propose a novel approach for fine-tuning three preeminent LLMs: MAmmoTH, LLaMA-2, and WizardMath using our MathQuest dataset.**

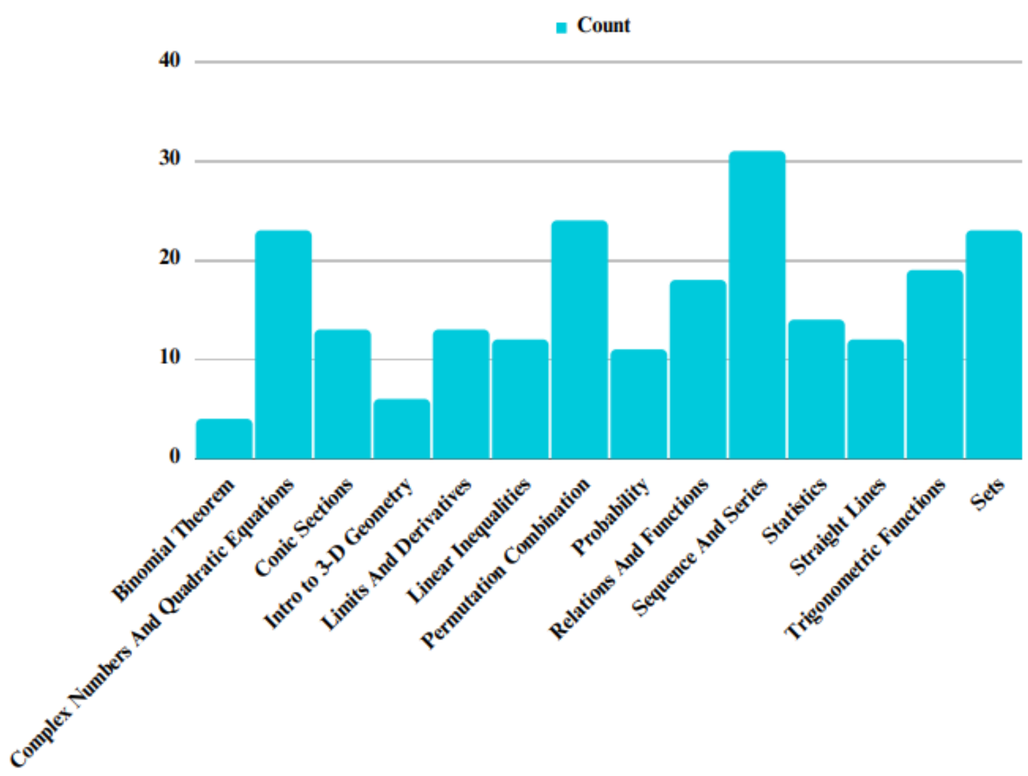


Our study not only assesses the performance of fine-tuned models on our "MathQuest" dataset but also their ability on other math reasoning datasets. Results show MAmmoTH-13B excels above others in solving mathematical problems, proving a reliable benchmark for NCERT math challenges.

MATHQUEST

- **We have meticulously curated our own dataset, referred to as MathQuest, sourcing problems from high school mathematics NCERT books.**
- **Our dataset comprises a total of 14 overarching mathematical domains, including sets, trigonometry, binomial theorem, and more.**
- **Our dataset contains total of 223 samples.**

Type	Range	Decimal Places (1 - 4)	Variables	Count
Small Integer	[-20, 20]	×	(x, y)	65,000
Small Decimal	[-20, 20]	✓	(x, y)	35,000
Small Decimal + Integer	[-20, 20]	✓	(x, y)	39,000
Large Integer	[-1000, 1000]	×	(x, y)	39,000
Large Decimal	[-1000, 1000]	✓	(x, y)	25,000
Large Decimal + Integer	[-1000, 1000]	✓	(x, y)	25,000
3 Terms	[-100, 100]	✓	(x, y, z)	25,000
4 Terms	[-100, 100]	✓	(w, x, y, z)	49,000
Total	-	-	-	302,000



- **We performed these experiments on both the 7B and 13B variants of three large language models (LLMs), i.e. LLaMA-2, WizardMath, and MammoTH.**
- **Our experiments involved two stages. In the first stage, we directly loaded the original model weights and performed inference on our designated test set. In the second stage, we fine-tuned these models using the Math-401 dataset.**

Model	# of Params	Accuracy					
		GSM-8K	DeepMind	NumGLUE	SimulEq	Math-401*	MathQuest
LLaMA-2	7B	30.0	46.0	45.0	15.0	17.0	10.6
LLaMA-2	13B	42.0	51.0	54.0	16.0	24.0	20.3
WizardMath	7B	64.0	55.0	52.0	29.0	15.0	16.01
WizardMath	13B	68.0	56.0	70.0	38.0	10.0	20.1
MAmmoTH	7B	56.0	50.0	62.0	24.0	16.0	18.5
MAmmoTH	13B	67.0	51.0	64.0	34.0	18.0	**24.0**

Table 3: Exact Match Accuracy Results on the set of 100 samples of 5 datasets and our dataset MathQuest **After** fine-tuning on Math-401 dataset. (*) refers to the set of Math-401 we augmented for fine-tuning.

Model	# of Params	Accuracy					
		GSM-8K	DeepMind	NumGLUE	SimulEq	Math-401*	MathQuest
LLaMA-2	7B	16.0	46.0	37.0	11.0	10.0	10.4
LLaMA-2	13B	22.0	50.0	42.0	15.0	10.0	14.1
WizardMath	7B	61.0	51.0	54.0	27.0	6.0	14.6
WizardMath	13B	65.0	55.0	70.0	36.0	8.0	14.3
MAmmoTH	7B	43.0	49.0	54.0	23.0	11.0	12.2
MAmmoTH	13B	44.0	48.0	56.0	26.0	14.0	18.1

Table 2: Exact Match Accuracy results on the set of 100 samples of 5 datasets and our dataset MathQuest **Before** fine-tuning on Math-401 dataset. (*) refers to the set of Math-401 we augmented for fine-tuning.

- **Key findings from Table 2 and Table 3 indicate that the top-performing model for our MathQuest dataset, following fine-tuning, is MAmmoTH 13B, achieving the highest accuracy at 24.0%.**
- **It is noteworthy that both MAmmoTH 7B and 13B produced outputs with precision up to two decimal places, highlighting their accuracy.**
- **Table 3 further reveals that MathQuest presents a greater challenge due to its complexity and diversity, leading to lower accuracy compared to other datasets.**

CONCLUSION

- **In summary, our approach enhances Large Language Models (LLMs) in acquiring vital reasoning skills for precise mathematical problem-solving. We introduce tailored question-answer pairs in our MathQuest dataset, encompassing single or multiple mathematical operators and expressions. These supportive simple and complex problems guide the model toward incremental problem-solving.**
- **Our experiments reveal that among the three models, MAmmoTH-13B emerges as the most proficient, achieving the highest level of competence in solving the presented mathematical problems. Consequently, MAmmoTH-13B establishes itself as a robust and dependable benchmark for addressing NCERT mathematics problems.**

NEURAL INFORMATION PROCESSING SYSTEMS
Generative AI for Education (GAIED) Workshop