NEURAL INFORMATION PROCESSING SYSTEMS

Generative AI for Education (GAIED) Workshop

Mathify: Evaluating Large Language Models on Mathematical Problem Solving Tasks

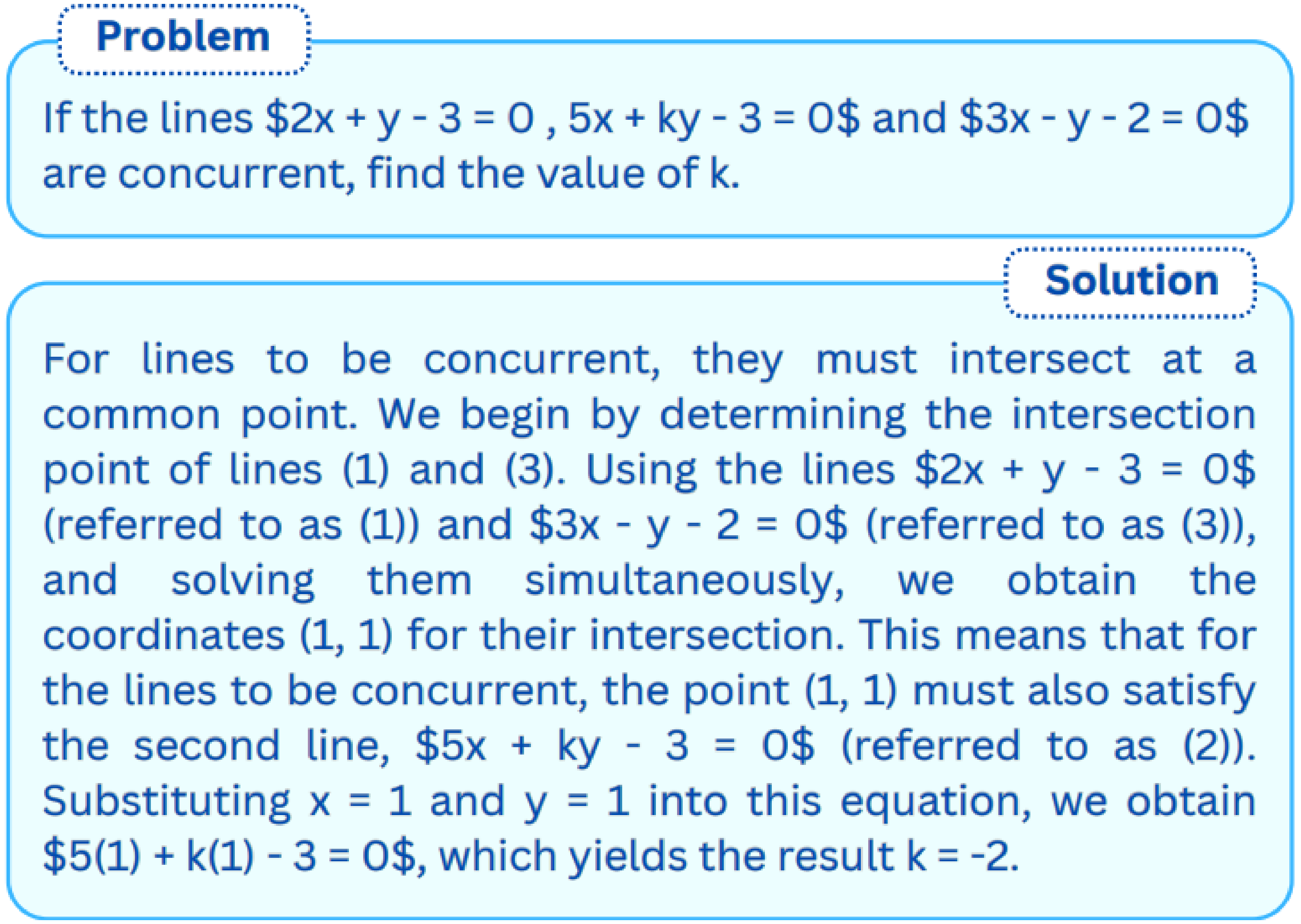
Avinash Anand, Mohit Gupta, Kritarth Prasad, Navya Singla, Sanjana Sanjeev, Jatin Kumar, Adarsh Raj Shivam, Rajiv Ratn Shah

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI (IIIT-DELHI)

BACKGROUND

- LLMs, grounded in the transformer architecture, have the capacity to glean long-range dependencies and contextual representations from vast corpora of text data.

- Previous approaches include building on the CoT framework, assesses multiple potential reasoning paths and selects answers via majority vote. Some use recent LLMs like GPT-3.5 to generate an output, provide feedback, and prompt the model for improvements.



Problem

If the lines $2x + y - 3 = 0 , 5x + ky - 3 = 0$ and $3x - y - 2 = 0$ are concurrent, find the value of k.

Solution

For lines to be concurrent, they must intersect at a common point. We begin by determining the intersection point of lines (1) and (3). Using the lines $2x + y - 3 = 0$ (referred to as (1)) and $3x - y - 2 = 0$ (referred to as (3)), and solving them simultaneously, we obtain the coordinates (1, 1) for their intersection. This means that for the lines to be concurrent, the point (1, 1) must also satisfy the second line, $5x + ky - 3 = 0$ (referred to as (2)). Substituting x = 1 and y = 1 into this equation, we obtain $5(1) + k(1) - 3 = 0$, which yields the result k = -2.

MOTIVATION

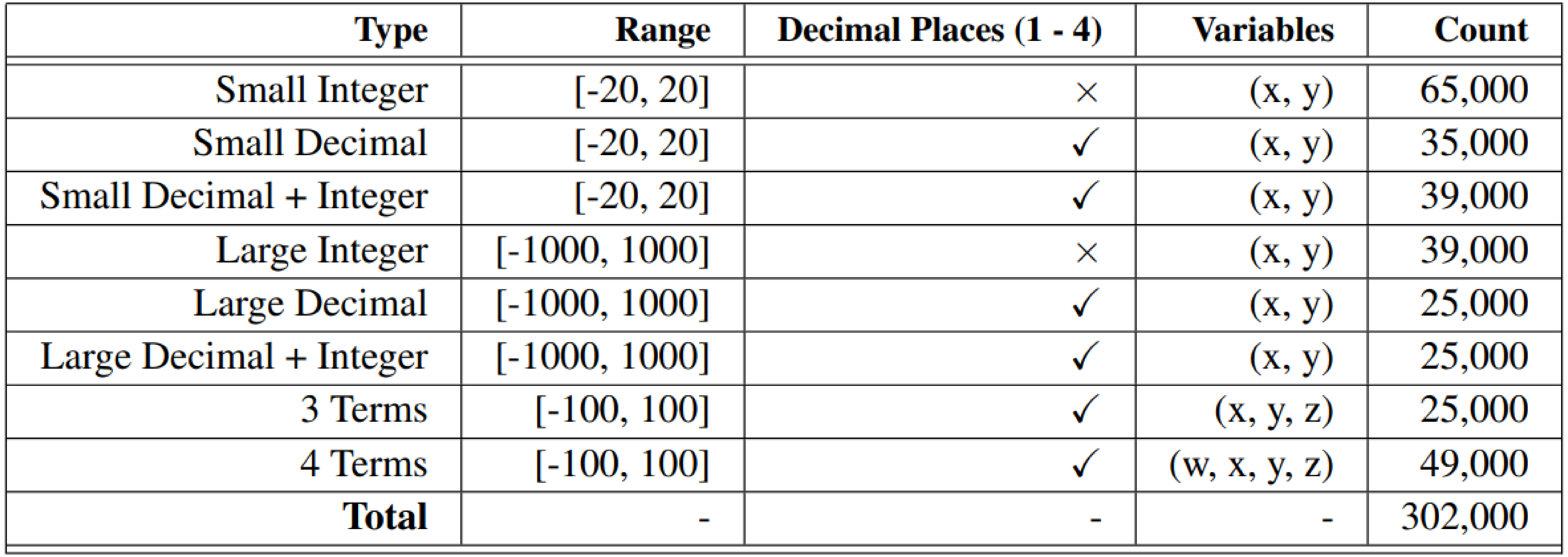
This research aims to enhance the mathematical problem-solving capabilities of large language models. Initially, we observed that existing open-source models such as LLaMA-2 struggled with elementary mathematical tasks like simple addition and subtraction. This observation served as the catalyst for our research, motivating us to improve LLMs' proficiency in comprehending and accurately solving mathematical problems.

Towards this,

- We introduce an extensive mathematics dataset called "MathQuest" sourced from the 11th and 12th standard Mathematics NCERT textbooks. This dataset encompasses mathematical challenges of varying complexity and covers a wide range of mathematical concepts.
- We conduct fine-tuning experiments with three prominent LLMs: LLaMA2, WizardMath, and MAmmoTH. These fine-tuned models serve as benchmarks for evaluating their performance on our dataset.

APPROACH

We employed the Math-401 dataset, which encompasses 401 samples of mathematical problems.



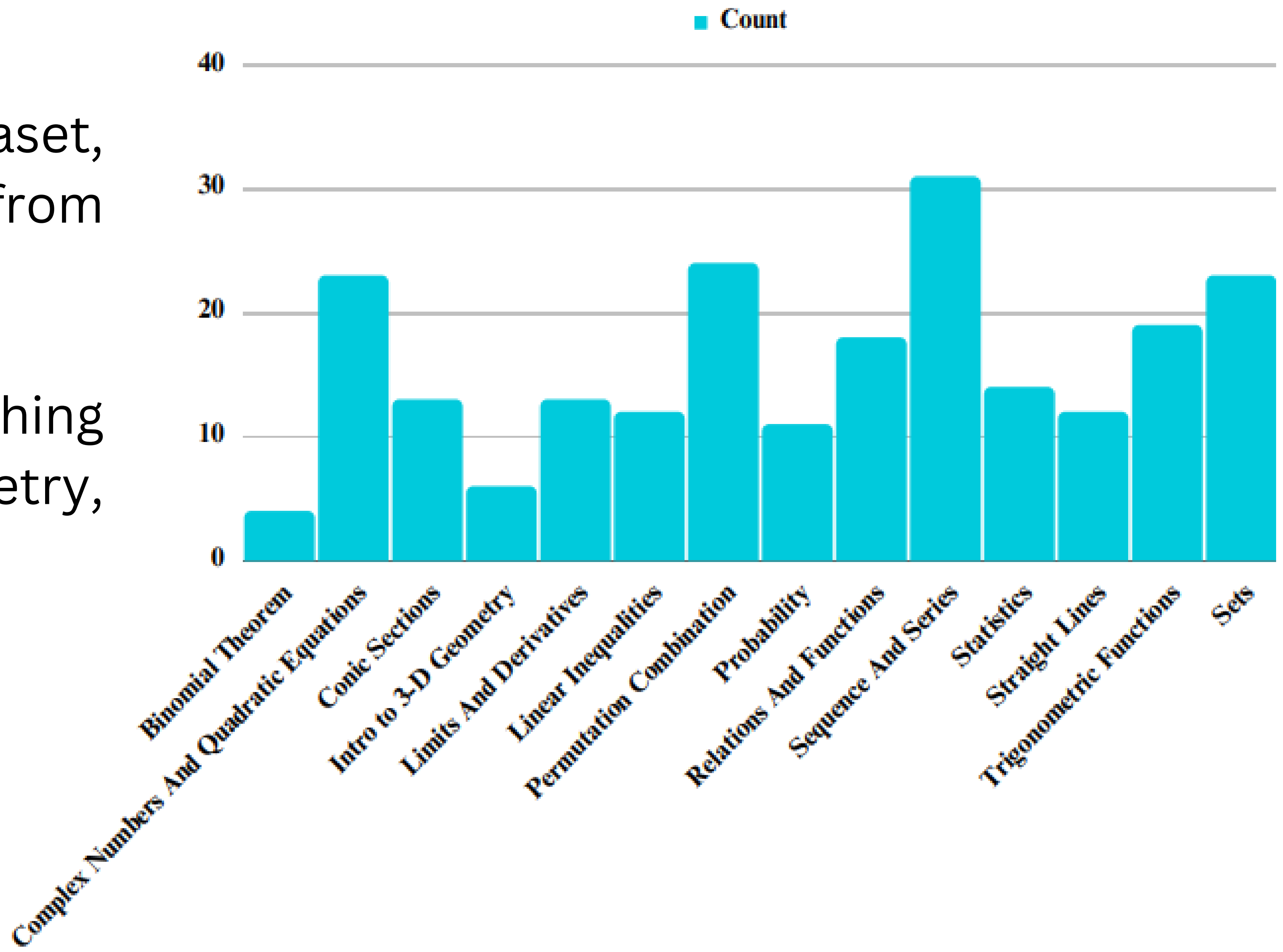
Type	Range	Decimal Places (1 - 4)	Variables	Count
Small Integer	[-20, 20]	×	(x, y)	65,000
Small Decimal	[-20, 20]	✓	(x, y)	35,000
Small Decimal + Integer	[-20, 20]	✓	(x, y)	39,000
Large Integer	[-1000, 1000]	×	(x, y)	39,000
Large Decimal	[-1000, 1000]	✓	(x, y)	25,000
Large Decimal + Integer	[-1000, 1000]	✓	(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

Table 1: The distribution of types of question in our augmented Math-401 dataset

APPROACH (Dataset)

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.



APPROACH (Dataset)

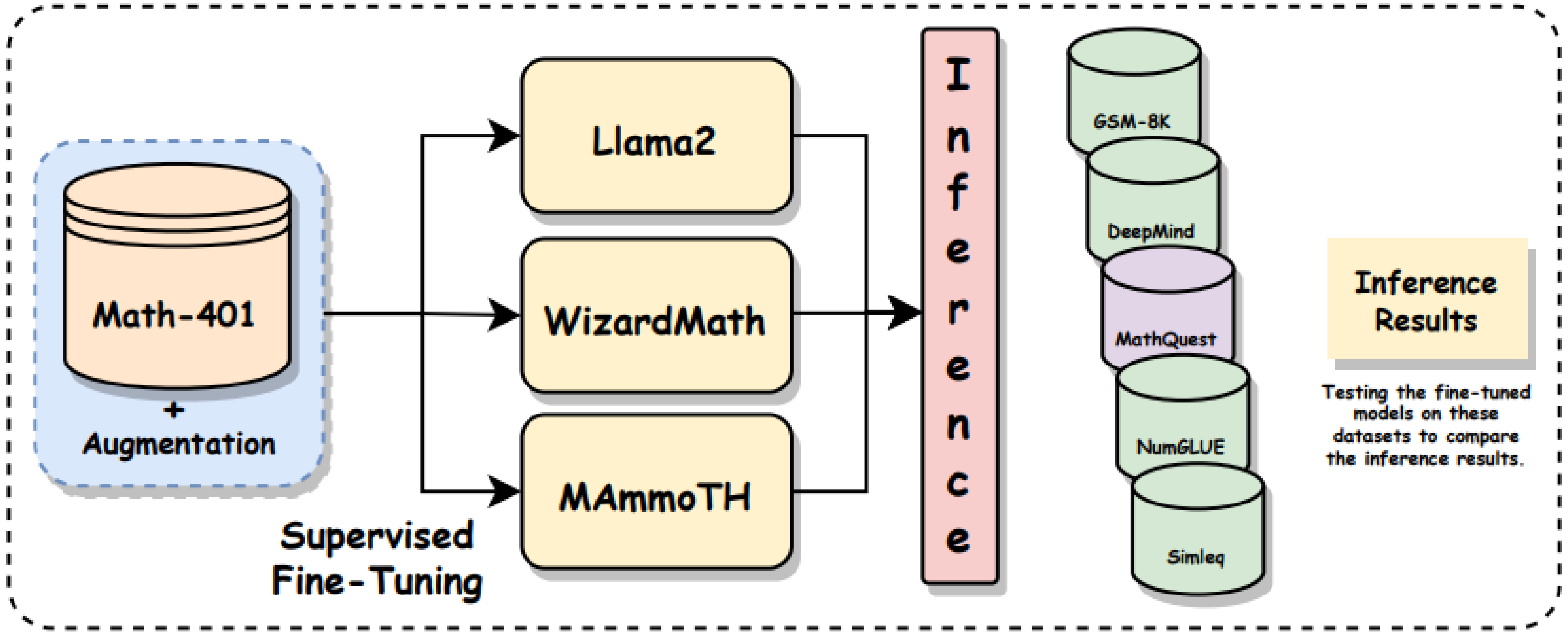


Figure 1: This figure shows the fine-tuning flow, the LLMs we use for fine-tuning, and the datasets we use for inference.

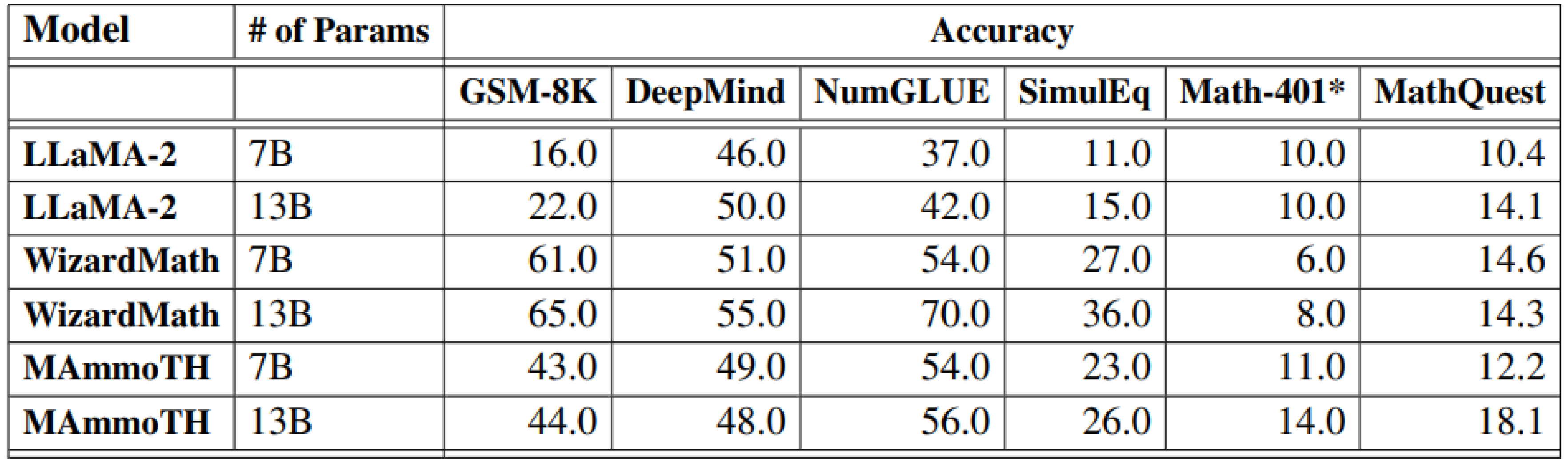
- We performed these experiments on both the 7B and 13B variants of three large language models (LLMs), i.e. LLaMA-2, WizardMath, and MammoTH.
- Our experiments were executed in two stages. In the first stage, we directly loaded the original model weights and carried out inference on our designated test set. In the second stage, we undertook the fine-tuning of these models using the Math-401 [36] dataset as a crucial step in the process.

APPROACH (Fine-Tuning)

RESULTS

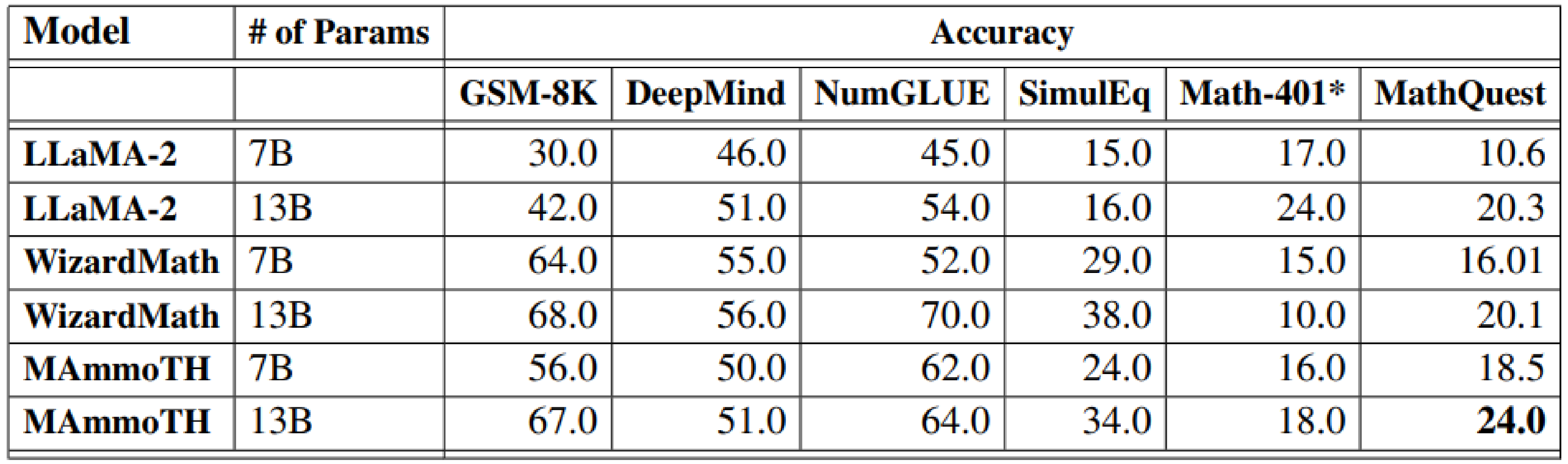
Our study encompasses evaluations conducted on our proprietary dataset, MathQuest, as well as five other publicly available datasets.

We organize our results into two distinct setups: before fine-tuning and after fine-tuning the models, with the primary aim of evaluating the model's learning capabilities.



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.



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.

The best-performing model is MAmmoTH13B for our dataset MathQuest, exhibiting the highest accuracy among all models after fine-tuning.

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.

CONCLUSION

THANK YOU

Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
Multimodal Digital Media Analysis (MIDAS) Lab