

# Large Language Models for Mathematical Reasoning: Progresses and Challenges

Presented by: Nevin Ada Cakmak and Katherine Lambert

# The paper aims to:

- explore ***math problems*** and data sets that have been investigated so far
- analyze ***different techniques*** used by LLMs for mathematical problem-solving
- factors that ***influence LLM performance*** in mathematical challenges
- examine persisting ***problems/challenges*** in the field

reasoning, geometry reasoning, logical reasoning, numeric common sense, scientific reasoning, and statistical reasoning. In addition, fine-grained metadata are available, including question type, answer type, language, source, category, task, grade level, and visual context.

## 4 Methodologies

We summarize these methods into three progressive levels: i) Prompting frozen LLMs, ii) Strategies enhancing frozen LLMs, and iii) Fine-tuning LLMs.

### 4.1 Prompting frozen LLMs

We organize prior work by typical LLMs.

**GPT-3.** Zong and Krishnamachari (2023) evaluated the use of GPT-3, a 175B parameter transformer model for three related challenges pertaining to math word problems: i) classifying word problems, ii) extracting equations from word problems, and iii) generating word problems.

**ChatGPT.** Shakarian et al. (2023) reported the first independent evaluation of ChatGPT on MWP, and found that ChatGPT's performance changes dramatically based on the requirement to show its work. Cheng and Zhang (2023) assessed ChatGPT, OpenAI's latest conversational chatbot and LLM, on its performance in elementary-grade arithmetic and logic problems, and found that ChatGPT performed better than previous models such as InstructGPT (Ouyang et al., 2022) and Minerva (Lewkowycz et al., 2022).

**GPT-4.** Wu et al. (2023) adapted and evaluated several existing prompting methods to the usage of GPT-4, including a vanilla prompt, Program-of-Thoughts prompt (Chen et al., 2023a), and Program Synthesis prompt (Drori et al., 2022). The study by Gu (2023) investigated the capability of GPT-4 to actively engage in math-oriented brainstorming sessions. This includes tasks like identifying new research problems, refining problem formulations, and suggesting potential methods or unconventional solutions, all achieved through iterative ideation with a human partner—a common practice in collaborative brainstorming with other professionals.

**GPT4V & Bard.** Lu et al. (2023a) presented MATHVISTA, a benchmark of evaluating mathematical reasoning in visual context, conducted a comprehensive, quantitative evaluation of three

LLMs (i.e., ChatGPT, GPT-4, Claude-2 (Bai et al., 2022)), two proprietary large multimodal models (LMMs) (i.e., GPT4V, Bard), and seven open-source LMMs, with Chain-of-Thought and Program-of-Thought.

**Multiple.** Wei et al. (2023) evaluated a variety of popular LLMs, including both commercial and open-source options, aiming to provide a benchmark tool for assessing the following question: to what grade level of Chinese elementary school math do the abilities of popular LLMs correspond?

### 4.2 Strategies enhancing frozen LLMs

**Preprocessing the math question.** An et al. (2023a) explored ChatGPT for the dataset SVAMP and observed that substituting numerical expressions with English expressions can elevate the performance.

**More advanced prompts.** Chain-of-thought (Wei et al., 2022), the first time to steer the LLMs to do **step-by-step math reasoning**, Self-Consistency (Wang et al., 2023) tried multiple Chain-of-Thought reasoning paths and leverage the **consistency** mechanism to discover a more probable answer. Zhou et al. (2023a) proposed a novel and effective prompting method, explicit code-based self-verification, to further boost the mathematical reasoning potential of GPT-4 Code Interpreter. This method employs a zero-shot prompt on GPT-4 Code Interpreter to encourage it to use code to **self-verify** its answers.

**Using external tool.** Yamauchi et al. (2023) employed an external tool, specifically the Python REPL, to correct errors in Chain-of-Thought. Their demonstration highlighted that integrating Chain-of-Thought and Python REPL using a markup language improves the reasoning capabilities of ChatGPT. In a related context, He-Yueya et al. (2023) introduced an approach that merges an LLM, Codex (Chen et al., 2021b), capable of progressively formalizing word problems into variables and equations, with an external symbolic solver adept at solving the generated equations. Program-of-Thought (Chen et al., 2023a) separates the computational aspect from the reasoning by utilizing a Language Model (primarily Codex) to articulate the reasoning procedure as a program. The actual computation is delegated to an external computer, responsible for executing the generated programs to arrive at the desired answer.

# LLM (Large Language Model)

*ChatGPT: Optimizing Language Models for Dialogue*

*We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it easy for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction precisely and provide a detailed response.*

TRY CHATGPT ➔

- a type of artificial intelligence (AI) program that has been fed enough data to be able to recognize and interpret human language or other types of complex data and ultimately generate text

(Cloudflare)

**Examples:** GPT 3, PaLM-2, ChatGPT-4, BingAI, Bard

# Related Work

---

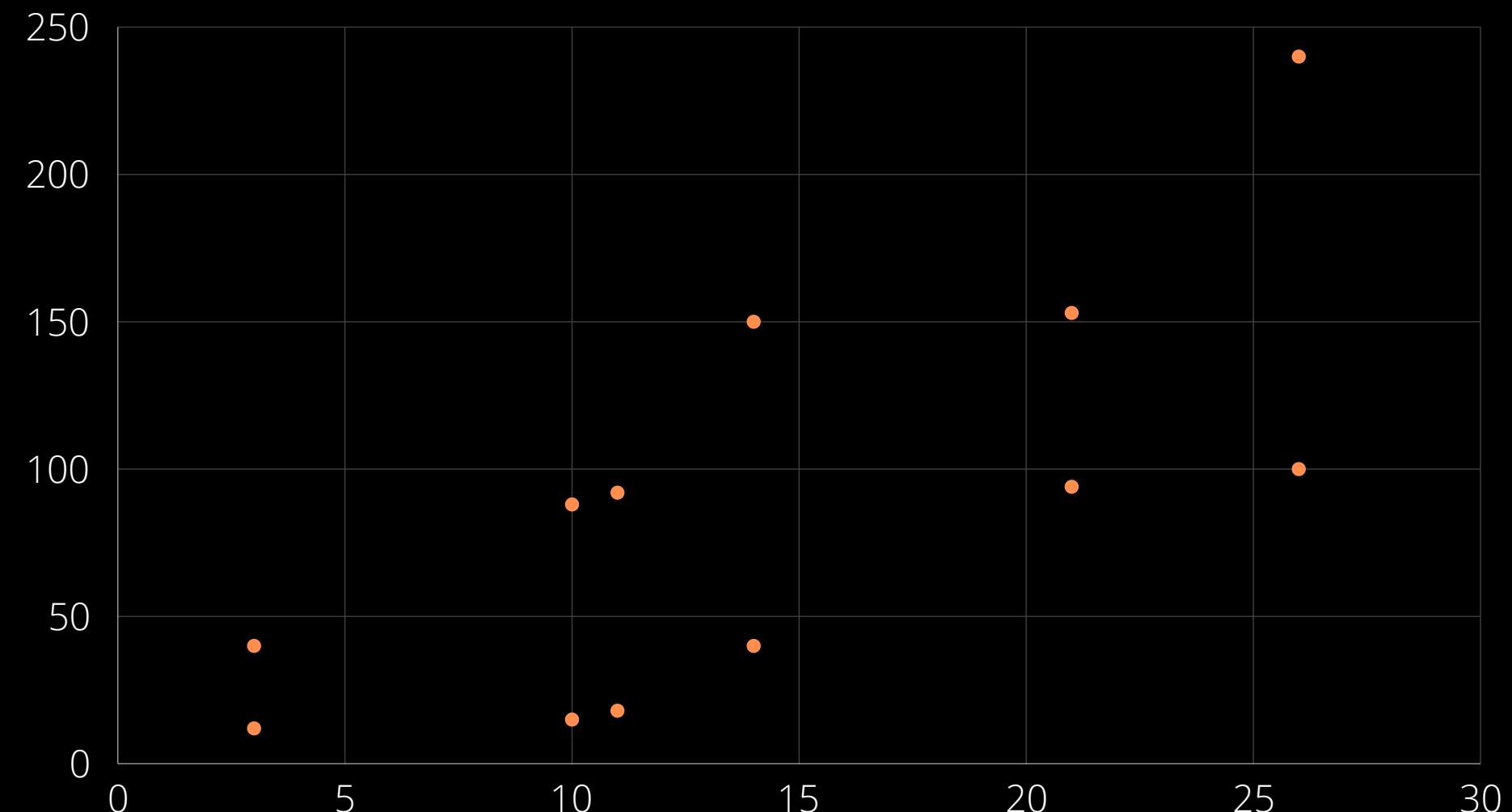
- Chang et al. (2023) conducted a comprehensive evaluation of LLMs, incorporating an examination of their performance in mathematical problem-solving.
- Testolin and Lu et al. delved into the application of Deep Learning in the domain of mathematical reasoning.
- In comparison with Liu et al., this paper paid more attention to various factors affecting model performance, took a broader perspective on the progress of LLM in the field of mathematics.

**And.... How does this paper contribute?**

# Math Problems and Datasets

# Math Problems and Datasets

- which problems were analysed as part of the training data
- sections:
  - arithmetic
  - word problems (MWP)
  - geometry
  - theorem proving
  - math in vision context





# Arithmetic

*“category of problems entails pure mathematical operations and numerical manipulation, devoid of the need for the model to interpret text, images, or other contextual elements”*

**q: 21 + 97**

**A: 118**

MATH-140 (Yuan et al., 2023)

# Math Word Problems (MWP)

*“mathematical exercises or scenarios presented in the form of written or verbal descriptions rather than straightforward equations in ARITHMETIC”*

problems require individuals to:

- decipher information provided
- identify relevant math concepts
- formulate equations or expressions to solve the given problem
- requires critical thinking

# Math Word Problems (MWP)

Q: Lily received \$20 from her mum. After spending \$10 on a storybook and \$2.5 on a lollipop, how much money does she have left?

A: \$ 7.5

(Wei et al., 2022)

# Math Word Problems (MWP)

	NAME	SIZE	LEVEL	NOTE
Q-A	CMATH (Wei et al., 2023) SAT-MATH (Zhong et al., 2023)	1.7K 220	E H	Chinese; grade 1-6 Multi-choice
Question-Equation-Answer	SVAMP (Patel et al., 2021)	1K	E	Three types of variations
	ASDIV (Miao et al., 2020)	2.3K	E	Problem type and grade level annotated
	MAWPS (Koncel-Kedziorski et al., 2016)	3.3K	E	Extension of ADDSUB, MULTIARITH, etc.
	PARAMAWPS (Raiyan et al., 2023)	16K	E	Paraphrased, adversarial MAWPS
	SINGLEEQ (Koncel-Kedziorski et al., 2015)	508	E	
	ADDSUB (Hosseini et al., 2014)	395	E	Only addition and subtraction
	MULTIARITH (Roy and Roth, 2015)	600	E	Multi-step reasoning
	DRAW-1K (Upadhyay and Chang, 2017)	1K	E	
	MATH23K (Wang et al., 2017)	23K	E	Chinese
	APE210K (Zhao et al., 2020)	210K	E	Chinese
Question-Rationale-Answer	K6 (Yang et al., 2023)	600	E	Chinese; grade 1-6
	CM17K (Qin et al., 2021)	17K	M H	Chinese; grade 6-12
	CARP (Zhang et al., 2023a)	4.9K	M	Chinese
	GSM8K (Cobbe et al., 2021)	8.5K	M	Linguistically diverse
	MATH (Hendrycks et al., 2021)	12.5K	H	Problems are put into difficulty levels 1-5
	PRM800K (Lightman et al., 2023)	12K	H	MATH w/ step-wise labels
	MATHQA (Amini et al., 2019)	37K	C	GRE examinations; have quality concern
	AQUA (Ling et al., 2017)	100K	C	GRE&GMAT questions
	ARB (Sawada et al., 2023)	105	C	Contest problems and university math proof
	GHOSTS (Frieder et al., 2023)	709	C	
	THEOREMQA-MATH (Chen et al., 2023b)	442	C	Theorem as rationale
	LILA (Mishra et al., 2022)	132K	H	Incorporates 20 existing datasets
	MATH-INSTRUCT (Yue et al., 2023)	260K	H	Instruction-following style
	TABMWP (Lu et al., 2023b)	38K	H	Tabular MWP; below the College level

**Table 1:** Datasets for Math Word Problems.

E = Elementary, M = Middle School, H = High School, C = College, H = Hybrid

## Question-Answer

*Q:* Lily received \$20 from her mum. After spending \$10 on a storybook and \$2.5 on a lollipop, how much money does she have left?

*A:* \$7.5

## Question-Rationale-Answer

*Q:* Beth bakes 4, or 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

*R:* Beth bakes 4 2 dozen batches of cookies for a total of  $4 * 2 = << 4 * 2 = 8 >>$  8 dozen cookies. There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of  $12 * 8 = << 12 * 8 = 96 >>$  96 cookies. She splits the 96 cookies equally amongst 16 people so they each eat  $96 / 16 = << 96 / 16 = 6 >>$  6 cookies.

*A:* 6

## Question-Equation-Answer

*Q:* Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?

*E:*  $8 - 3$

*A:* 5 (optional)

## Tabular MWP

BEADS	\$/KILOGRAM
heart-shaped	3
rectangular	2
spherical	2
oval	2

**Table 2:** Table for the tabular MWP example.

*T:* Table 2

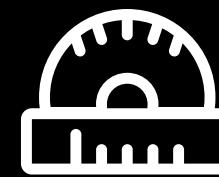
*Q:* Henrik bought 2.5 kilograms of oval beads. How much did he spend? (Unit: \$)

*A:* 5

# MWP Generation

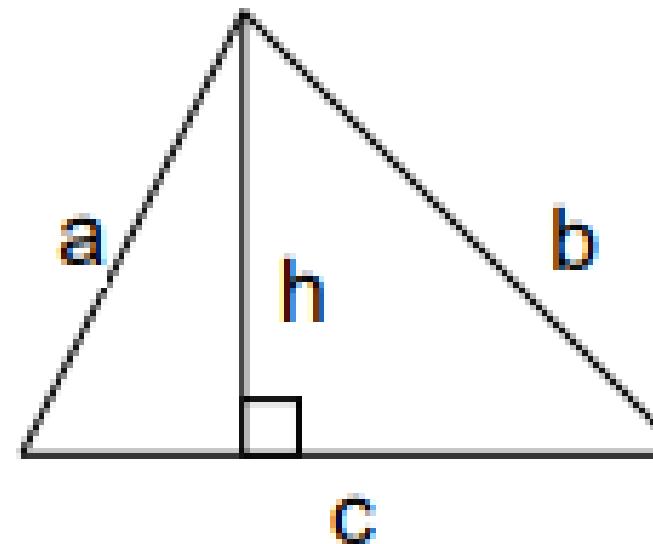
- focused on *generation* of math word problems
- found a significant drop in backward reasoning compared to forward reasoning

- Perceived Difficulty Level:
  - Modifications considered neutral or simple for LLMs could be perceived as challenging by human students
- Imperfect Logical Coherence:
  - logical coherence is not perfect in few cases. For example, the phrases like “25 hour in a day” and “37 days in a month”
- Cannot Alter Words:
  - Usually need to alter numerical values, but there are problems where words also represent numeric values (ex. "doubles" and "three")



# Geometry

- demands a spacial understanding of shapes, sizes, and their relationships
- involves applying geometric principles & theorems and formulas to analyze and deduce properties of geometric figures in question



**Q:**  $a=7$  inches;  $b=24$  inches;  $c=25$  inches;  
 $h=5.4$  inches; What is its area? (Unit: square inches)

**A:** 24.03



You

think of a right triangle.  $a = 24$ ,  $b = 7$ ,  $c=25$  and its height of  $c$  is 5.4 inches. what is its area (unit: square inches)



Copilot

Certainly! Let's find the area of the right triangle using the given information.

Given:

- Side **a** (base) = 24 inches
- Side **b** = 7 inches
- Hypotenuse **c** = 25 inches
- Height (perpendicular to base **c**) = 5.4 inches

We can use the formula for the area of a right triangle:

[ \text{Area} = \frac{1}{2} \cdot b \cdot h ]

where:

- (**b**) is the base (side **a**)
- (**h**) is the height (perpendicular to base **c**)

Substitute the given values:

[ \text{Area} = \frac{1}{2} \cdot 24 \cdot 5.4 ]

Calculating:

[ \text{Area} = 64.8 , \text{square inches} ]

Therefore, the area of the right triangle is approximately **64.8 square inches**. 🔍▲



You

think of a right triangle.  $a = 24$ ,  $b = 7$ ,  $c=25$  and its height of  $c$  is 5.4 inches. what is its area (unit: square inches)



ChatGPT

To find the area of a right triangle, you can use the formula:

$$\text{Area} = \frac{1}{2} \times \text{base} \times \text{height}$$

Given that  $a = 24$ ,  $b = 7$ , and  $c = 25$  and the height  $h_c = 5.4$ , you can use the fact that the base of the triangle is either  $a$  or  $b$ , as the height is perpendicular to the hypotenuse, and the hypotenuse is the longest side, which is  $c$ .

Using  $b$  as the base:

$$\text{Area} = \frac{1}{2} \times b \times h_c$$

$$\text{Area} = \frac{1}{2} \times 7 \times 5.4$$

$$\text{Area} = \frac{1}{2} \times 7 \times 5.4$$

$$\text{Area} = 18.9 \text{ square inches}$$

So, the area of the triangle is 18.9 square inches.



# Automated Theorem Proving

---



Automated theorem proving involves...



This seems almost too good! What about challenges?

# Math in Vision-Language Context

CHARTQA (Masry et al., 2022):

- Explores complex reasoning with 9.6K human-written and 23.1K model-generated questions on chart data.

MATHVISTA (Lu et al., 2023a):

- Contains 6K questions covering seven types of mathematical reasoning, with detailed metadata.

# Methodologies

1

Prompting frozen LLMs,

2

Strategies enhancing frozen LLMs

3

Fine-tuning LLMs.

# i) Prompting Frozen LLM's

Prior work to: GPT-3, ChatGPT, GPT-4, GPT4V & Bard...

- ChatGPT's performance changes dramatically based on the requirement to show its work
- MATHVISTA: a benchmark for evaluating mathematical reasoning in a visual context
- Chain-of-Thought, Program-of-Thought
- Evaluated transformer model for three related challenges: i) classifying word problems, ii) extracting equations from word problems, and iii) generating word problems.
- Investigated the capability of GPT-4 to actively engage in math-oriented brainstorming sessions, with a human partner

## ii) Strategies Enhancing Frozen LLMs

1

Preprocessing the math question:  
Substituting numerical  
expressions with English ones  
can elevate the performance.

2

Using more advanced prompts

3

Using an external tool

## ii) Strategies Enhancing Frozen LLMs

4

Improving the whole interaction

5

Considering more comprehensive factors in evaluation

- a) Confidence Provision
- b) Verifiable Explanations

$$\int_V \operatorname{div} F \, dV = \int_{\partial V} F \cdot d\Sigma$$

# Fine-tuning LLMs

## Generating intermediate steps:

- researchers tried to fine-tune LLMs by getting them to generate step-by-step solutions

## Learning to select in-context examples:

- study few-shot GPT-3 performance
- performance tends to be unstable
- new approach: usage of a tool PromptGPT
- finds optimal in-context examples from the candidate pool

# Fine-tuning LLMs

## Learning an answer verifier:

- researchers fine-tuned a GPT-3 model as a verifier
- observed improved correctness in LLMs with multiple attempts
- --> fine-tuned a PaLM-2 model as a solution generator, evaluator, and generator again

# Fine-tuning LLMs

## Learning from Enhanced Dataset:

- Imitated error-driven human learning process by fine-tuning LLMs with mistake-correction data
- training on variants of problem statements + implementing a voting mechanism = improved the mathematical reasoning

# Fine-tuning LLMs



## Teacher-Student knowledge distillation:

- explaining existing exercises -> identifying the learning needs
- resulting LLM had better accuracy than more advanced LLMs

## Finetuning on many datasets:

- explored the usage of composite datasets
- resulting models demonstrated unprecedented potential in cross-dataset generalization

# Analysis

## 1) LLMs's robustness in math

- (i) Adversarial examples crafted from better-performing LLMs can also harm less accurate LLMs
- (ii) Complex MWPs are more vulnerable to attack
- (iii) We can improve the robustness of LLMs by using adversarial samples in few-shot prompts

- For LLMs that aren't specifically tailored with instructions, larger models are often sensitive to changes in the correct answer for math problems but may not be more robust overall.
- GPT-3 models that have been tuned with instructions show improvements in both sensitivity and robustness. However, their robustness decreases as the complexity of the problems increases.

# Analysis

## 2) Factors in influencing LLMs in mathematics

### Model Scale

Larger models generally perform better, but a performance plateau is observed.

### Prompts

Without prompts, performance drops. Instruction tuning in pre-training also emerges as a significant factor

### Tokenization

Models lacking specialized tokenization for arithmetic are less effective than those with advanced methods.

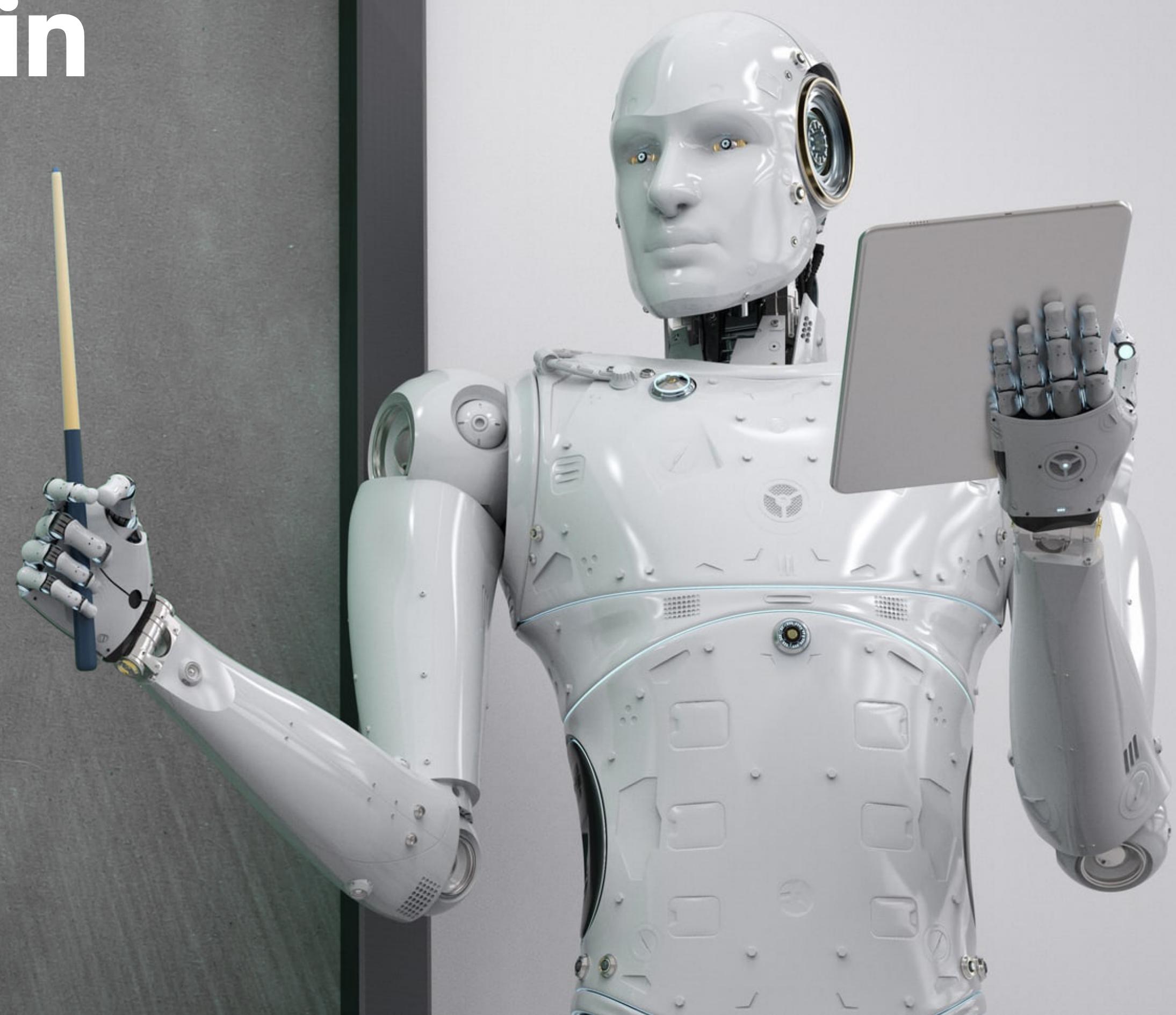
### Pre-training Corpus

Enhanced arithmetic skills in LLMs correlate with the inclusion of code and LATEX in pre-training data.

There is a distinction between arithmetic and reasoning skills.

# Perspectives in Pedagogy

**Pedagogy:** of, relating to, or befitting a teacher or education (Merriam-Webster)



# The benefits of using LLMs in education



- 01 Foster Critical Thinking and Problem-Solving
- 02 Hints provided are clear and detailed
- 03 Introduce a conversational style
- 04 Extending capabilities

# The disadvantages of using LLMs in education



- 01 Potential Misinterpretation of questions, leading to confusion
- 02 Limited understanding of Learning Styles
- 03 Privacy and Security Issues

# What are some other challenges faced by LLMs?

---

**Data-driven and limited generalization**

**Human-LLM interactions**

**Fragility in math reasoning**

# Conclusions



## What did this paper do?

This paper on LLMs for Mathematics delves into various aspects of LLMs in mathematical reasoning, including their capabilities and limitations.



## What did this paper do?

It highlights the advancements in LLMs, their application in educational settings, and the need for a human-centric approach in math education.



## Future Work

Researchers hope this paper will guide and inspire future research in the LLM community, fostering further advancements in diverse mathematical contexts.

# Our Review of the Paper...

- The paper lacks coherence due to its mixed topics and insufficient explanation for certain points.
- The method of paper selection remains unspecified.
- The conclusion, while brief, fails to effectively tie together all the presented information.
- Additionally, the paragraph of "Math in VL" does not contain enough information.
- It's noted that certain sections only reference one study, such as the "Fine-Tuning LLMS," which undermines the goal of summarizing multiple research findings.
- Furthermore, there's a tendency to list research without detailing the outcomes, leading to a disjointed conclusion. These issues highlight the need for concise and cohesive content in the presentation.