Large Language Models for Mathematical Reasoning: Progresses and Challenges

Janice Ahn♠ **Rishu Verma**♠ **Renze Lou**♠ **Di Liu**♢
Rui Zhang♠ and **Wenpeng Yin**♠
♠The Pennsylvania State University ♢ Temple University
{jfa5672, wenpeng}@psu.edu; diliu@temple.edu

Abstract

Mathematical reasoning serves as a cornerstone for assessing the fundamental cognitive capabilities of human intelligence. In recent times, there has been a notable surge in the development of Large Language Models (LLMs) geared towards the automated resolution of mathematical problems. However, the landscape of mathematical problem types is vast and varied, with LLM-oriented techniques undergoing evaluation across diverse datasets and settings. This diversity makes it challenging to discern the true advancements and obstacles within this burgeoning field. This survey endeavors to address four pivotal dimensions: i) a comprehensive exploration of the various mathematical problems and their corresponding datasets that have been investigated; ii) an examination of the spectrum of LLM-oriented techniques that have been proposed for mathematical problem-solving; iii) an overview of factors and concerns affecting LLMs in solving math; and iv) an elucidation of the persisting challenges within this domain. To the best of our knowledge, this survey stands as one of the first extensive examinations of the landscape of LLMs in the realm of mathematics, providing a holistic perspective on the current state, accomplishments, and future challenges in this rapidly evolving field.

1 Introduction

Mathematical reasoning is crucial to human intelligence, driving ongoing efforts in the AI community to autonomously tackle math challenges. This pursuit inherently calls for an augmentation of AI capabilities, delving into the intricate realms of textual comprehension, image interpretation, tabular analysis, symbolic manipulation, operational logic, and a nuanced grasp of world knowledge. As the AI landscape evolves, the endeavor to empower machines with a comprehensive understanding of diverse mathematical facets becomes not only a testament to technological prowess but also a pivotal stride towards achieving a more generalized and adept AI.

In recent times, the landscape of AI has been reshaped by the ascendancy of Large Language Models (LLMs) as formidable tools for automating intricate tasks. Notably, LLMs have proven to be potent assets in unraveling the nuances of mathematical problem-solving (Romera-Paredes et al., 2023; Imani et al., 2023). Their language capabilities fuel focused exploration in utilizing them for mathematical reasoning, uncovering fresh insights into the synergy between language and logic.

However, amid this progress, the current state of LLM-oriented research in mathematics presents a complex panorama. Diverse mathematical problem types pose a formidable challenge, exacerbated by the varied evaluation metrics, datasets, and settings employed in the assessment of LLM-oriented techniques (Testolin, 2023; Lu et al., 2023c). The lack of a unified framework hampers our ability to gauge the true extent of progress achieved and impedes a coherent understanding of the challenges that persist in this evolving field.

This survey endeavors to cast a spotlight on the multifaceted landscape of LLMs in the realm of mathematics. We plan to traverse four crucial dimensions: a meticulous exploration of math problem types and the datasets associated with them; an in-depth analysis of the evolving techniques employed by LLMs in mathematical problem-solving; an examination of factors that affect the LLMs solving math problems; and a critical discussion on the persisting challenges that loom over this burgeoning field.

To our knowledge, this survey marks one of the first comprehensive examinations of LLMs specifically tailored for mathematics. By weaving together insights from various dimensions, we aim to provide a holistic understanding of the current state of affairs in LLM-driven mathematical reasoning, shedding light on achievements, challenges, and the uncharted territories that await exploration in this captivating intersection of language and logic.

2 Related Work

To the best of our knowledge, the existing literature on summarizing mathematical research, particularly within the context of LLMs, remains limited. Notably, Frieder et al. (2023a) compared two ChatGPT versions (9-January-2023 and 30-January-2023) and GPT-4 for four math-related problems: producing proofs, filling holes in proofs, acting as a mathematical search engine and computation. More importantly, they summarized some insightful strategies regarding how LLMs can help mathematicians and advocated a more collaborative approach, incorporating human expertise and LLM automation, for theorem proving. Chang et al. (2023) conducted a comprehensive evaluation of LLMs, incorporating an examination of their performance in mathematical problem-solving, albeit with a relatively brief exploration of the mathematical field. Conversely, both (Testolin, 2023) and (Lu et al., 2023c) delved into the application of Deep Learning in the domain of mathematical reasoning. Our work distinguishes itself on three fronts: firstly, we concentrate on LLMs, providing a more in-depth analysis of their various advancements; secondly, beyond merely reporting progress, we engage in a thorough discussion of the challenges inherent in this trajectory; and thirdly, we extend our scrutiny to encompass the perspective of mathematics pedagogy. In doing so, we contribute a nuanced perspective that seeks to broaden the understanding of LLMs in the context of mathematical research.

The only work contemporaneous with ours is (Liu et al., 2023b). In comparison, our contribution lies in: i) not only introducing various methods but also paying more attention to various factors affecting model performance; ii) taking a broader perspective on the progress of LLM in the field of mathematics, elucidating not only from the AI perspective but also from the perspective of education. It emphasizes that the pursuit of model performance alone, while neglecting human factors, is something that needs attention.

3 Math Problems & Datasets

This section concisely overviews prominent mathematical problem types and associated datasets, spanning ARITHMETIC, MATH WORD PROBLEMS, GEOMETRY, AUTOMATED THEOREM PROVING, and MATH IN VISION CONTEXT.

3.1 Arithmetic

This 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. An illustrative example is presented below, where "$\mathcal{Q}$" denotes questions and "$\mathcal{A}$" for answers.

$\mathcal{Q}$: `21 + 97`
$\mathcal{A}$: `118`

The dataset MATH-140 (Yuan et al., 2023) contains 401 arithmetic expressions for 17 groups.

3.2 Math Word Problems

MATH WORD PROBLEMS (MWP) are mathematical exercises or scenarios presented in the form of written or verbal descriptions rather than straightforward equations in ARITHMETIC. These problems require individuals to decipher the information provided, identify relevant mathematical concepts, and formulate equations or expressions to solve the given problem. MWP often reflect real-world situations, allowing individuals to apply mathematical principles to practical contexts. Solving these problems typically involves critical thinking, problem-solving skills, and the application of mathematical operations to find a solution.

MWP invariably comprise a question ($\mathcal{Q}$) and its corresponding final answer ($\mathcal{A}$) (referred to as *Question-Answer*). However, the presence or absence of additional clues can give rise to various versions of these problems. Variations may emerge based on factors such as the availability of an equation ($\mathcal{E}$; referred to as *Question-Equation-Answer*) or the provision of a step-by-step rationale ($\mathcal{R}$; *Question-Rationale-Answer*) to guide the problem-solving process.

Question-Answer. The instance of this type of MWP consists of a question ($\mathcal{Q}$) and the final answer ($\mathcal{A}$), such as:

$\mathcal{Q}$: `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?`
$\mathcal{A}$: `$7.5`

	NAME	SIZE	LEVEL	NOTE
Q-A	CMATH (Wei et al., 2023)	1.7K	E	Chinese; grade 1-6
	SAT-MATH (Zhong et al., 2023)	220	H	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
	K6 (Yang et al., 2023)	600	E	Chinese; grade 1-6
	CM17K (Qin et al., 2021)	17K	M H	Chinese; grade 6-12
Question-Rationale-Answer	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., 2023b)	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-Equation-Answer. Compared with *Question-Answer*, this MWP type provides the equation solution, such as

$\mathcal{Q}$: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?
$\mathcal{E}$: $8-3$
$\mathcal{A}$: 5 (optional)

Question-Rationale-Answer. This type of MWP includes answers and reasoning paths, akin to the Chain-of-Thought method, which explicates reasoning steps rather than defining problem types (Wei et al., 2022). The rationale guides correct problem-solving and serves as a valuable reference for model training, including fine-tuning and few-shot learning.

$\mathcal{Q}$: 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?
$\mathcal{R}$: 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.
$\mathcal{A}$: 6

Table 1 lists most datasets that are summarized in three categories: *Question-Answer*, *Question-Equation-Answer*, and *Question-Rationale-Answer*. In addition to the above three MWP types of conventional styles, recent work studied MWP in given tables and even MWP generation.

Tabular MWP. TabMWP (Lu et al., 2023b) is the first dataset to study MWP over tabular context on open domains and is the largest in terms of data size. Each problem in TabMWP is accompanied by a tabular context, which is represented in three formats: an image, a semi-structured text, and a structured table.

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

Table 2: Table for the tabular MWP example.

$\mathcal{T}$: Table 2
$\mathcal{Q}$: `Henrik bought 2.5 kilograms of oval beads. How much did he spend? (Unit: $)`
$\mathcal{A}$: `5`

MWP Generation. Instead of deriving the answer for a given math question, this type of mathematical reasoning tries to generate MWP questions. For example, Wang et al. (2021) fine-tuned GPT-2 (Radford et al., 2019) on equation-to-MWP instances for MWP generation. The effectiveness of GPT-3's question-generation capabilities was assessed by Zong and Krishnamachari (2023), who instructed the model to generate a question similar to a provided MWP question. Deb et al. (2023) analyzed a group of LLMs (GPT-4, GPT-3.5, PaLM-2 (Anil et al., 2023), and LLaMa (Touvron et al., 2023a)), and found a significant drop in accuracy for backward reasoning compared to forward reasoning. Norberg et al. (2023) used GPT-4 to rewrite human-written MWP, reporting optimal readability, lexical diversity, and cohesion scores, although GPT-4 rewrites incorporated more low-frequency words.

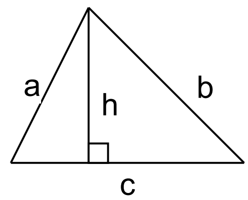
3.3 Geometry

Compared with MWP, GEOMETRY problems involve a distinct set of challenges. While MWP often requires logical reasoning and arithmetic operations, geometry problems demand a spatial understanding of shapes, sizes, and their interrelationships. Solving geometry problems typically entails applying geometric principles, theorems, and formulas to analyze and deduce properties of geometric figures. Furthermore, current geometry approaches mainly rely on symbolic methods and

Name	Size
GeoShader (Alvin et al., 2017)	102
GeoS (Seo et al., 2015)	186
GeoS++ (Sachan et al., 2017)	1.4K
GeoS-OS (Sachan and Xing, 2017)	2.2K
Geometry3K (Lu et al., 2021)	3K
GeoQA (Chen et al., 2021a)	5K
UniGeo (Chen et al., 2022)	14.5K

Table 3: Geometry datasets

predefined search heuristics, highlighting the specialized strategies required in this domain (Trinh et al., 2024). This contrast in problem-solving approaches highlights the multifaceted nature of mathematical challenges and the varied skill sets required in different mathematical domains. An example can be seen as follows and Table 3 lists mainstream datasets.



$\mathcal{Q}$: `a=7 inches; b=24 inches; c=25 inches; h=5.4 inches; What is its area? (Unit: square inches)`
$\mathcal{A}$: `24.03`

3.4 Automated theorem proving

In the specialized area of Automated Theorem Proving (ATP), the inherent challenges are unique and encompass a wide spectrum, akin to those found in distinct mathematical fields. ATP's core focus is on autonomously constructing proofs for specified conjectures, requiring a blend of logical analysis and a profound grasp of formal languages, supported by an extensive knowledge base. Its application is crucial in areas like the validation and development of both software and hardware systems.

For example, the MiniF2F dataset (Zheng et al., 2022) stands out in ATP, featuring a series of complex Olympiad-level mathematical problems, designed to evaluate theorem-proving systems including Metamath (Yu et al., 2023), Lean (Han et al., 2022), and Isabelle (Wenzel et al., 2008). In a similar vein, the HOList benchmark (Bansal et al., 2019), with its comprehensive array of theorem statements from various corpora, sets a sequential proving challenge for ATP systems, where each theorem must be proved using only the lemmas preceding it. Additionally, the CoqGym dataset (Yang and Deng, 2019) provides a broad ATP environment, showcasing a rich collection of more than 71,000 proofs penned by humans, all within the framework of the Coq proof assistant. These datasets illustrate the diverse methodologies and skillsets necessary in ATP, reflecting the multi-faceted nature of solving mathematical problems.

3.5 Math in vision-language context

ChartQA (Masry et al., 2022), with 9.6K human-written questions and 23.1K model-generated questions have explored a variety of complex reasoning questions that involve several logical and arithmetic operations over charts. MathVista (Lu et al., 2023a): size: 6K; it features seven types of mathematical reasoning: algebraic reasoning, arithmetic 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.

Improving the whole interaction. Wu et al. (2023) introduced MathChat, a conversational framework designed for chat-based LLMs. In this framework, math problems from the MATH dataset are resolved through a simulated conversation between the model and a user proxy agent.

Considering more comprehensive factors in evaluation. While accuracy is crucial in evaluating LLMs for math problem-solving, it shouldn't be the sole metric. Other important dimensions include: i) **Confidence Provision:** Imani et al. (2023)'s "MathPromper" boosts LLM performance and confidence by generating algebraic expressions, providing diverse prompts, and evaluating consensus among multiple runs. ii) **Verifiable Explanations:** Gaur and Saunshi (2023) used concise, verifiable explanations to assess LLM reasoning, revealing their proficiency in zero-shot solving of symbolic MWPand their ability to produce succinct explanations.

4.3 Fine-tuning LLMs

Learning to select in-context examples. As indicated by prior research, few-shot GPT-3's performance is susceptible to instability and may decline to near chance levels due to the reliance on in-context examples. This instability becomes more pronounced when dealing with intricate problems such as TABMWP. In addressing this issue, Lu et al. (2023b) introduced PROMPTPG, which can autonomously learn to select effective in-context examples through policy gradient interactions with the GPT-3 API, eliminating the need for manually designed heuristics.

Generating intermediate steps. Nye et al. (2021) initiated the fine-tuning of decoder-only LLMs, ranging from 2M to 137B in size. Their approach involved training these models to solve integer addition and polynomial evaluation by generating intermediate computation steps into a designated "scratchpad." In a related effort, Zhang et al. (2023b) introduced a fine-tuning strategy for GPT-2 or T5, enabling them to produce step-by-step solutions with a combination of textual and mathematical tokens leading to the final answer. Additionally, Yang et al. (2023) applied a step-by-step strategy in fine-tuning a series of GLM models (Zeng et al., 2023), specifically tailored for solving distinct Chinese mathematical problems. Minerva, developed by Lewkowycz et al. (2022), enhances LLMs' ability to generate intermediate steps in complex math problems. Its fine-tuning of diverse datasets enables nuanced, step-by-step problem-solving, demonstrating advanced handling of intricate mathematical concepts.

Learning an answer verifier. OpenAI researchers, per Cobbe et al. (2021), fine-tuned a GPT-3 model of 175B as a verifier, assigning probabilities to solution candidates. In exploring reexamination processes for MWP solving, Bin et al. (2023) introduced Pseudo-Dual Learning, involving solving and reexamining modules. For MWP solution, Zhu et al. (2023) developed a cooperative reasoning-induced PLM, with GPT-J (Wang and Komatsuzaki, 2021) generating paths and DeBERTa-large (He et al., 2021) supervising evaluation. Google researchers, as per Liu et al. (2023c), observed improved correctness in LLMs with multiple attempts, which hints that LLMs might generate correct solutions while struggling to differentiate between accurate and inaccurate ones. They sequentially fine-tuned their PaLM 2 model (Anil et al., 2023) as a solution generator, evaluator, and generator again.

Learning from enhanced dataset. Emulating the error-driven learning process observed in human learning, An et al. (2023b) conducted fine-tuning on various open-source LLMs within the LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b), CodeLLaMA (Rozière et al., 2023), WizardMath (Luo et al., 2023), MetaMath (Yu et al., 2023), and Llemma (Azerbayev et al.,

2023) families. This fine-tuning utilized mistake-correction data pairs generated by GPT-4. To mitigate over-reliance on knowledge distillation from LLM teachers, Liang et al. (2023a) fine-tuned LLaMA-7B on existing mathematical problem datasets that exhibit diverse annotation styles. In a related approach, Raiyan et al. (2023) demonstrated that training on linguistic variants of problem statements and implementing a voting mechanism for candidate predictions enhance the mathematical reasoning and overall robustness of the model.

Teacher-Student knowledge distillation. Liang et al. (2023b) utilized GPT-3 to coach a more efficient MWP solver (RoBERTa-based encoder-decoder (Liu et al., 2019)). They shifted the focus from explaining existing exercises to identifying the student model's learning needs and generating new, tailored exercises. The resulting smaller LLM achieves competitive accuracy on the SVAMP dataset with significantly fewer parameters compared to state-of-the-art LLMs.

Finetuning on many datasets. Mishra et al. (2022) conducted fine-tuning on a series of GPT-Neo2.7B causal language models (Black et al., 2021) using LILA, a composite of 20 existing math datasets. Similarly, Yue et al. (2023) created "MathInstruct", a meticulously curated instruction tuning dataset. Comprising 13 math datasets with intermediate Chain-of-Thought and Program-of-Thought rationales, this dataset was used to fine-tune Llama (Touvron et al., 2023a,b; Rozière et al., 2023) models across different scales. The resulting models demonstrate unprecedented potential in cross-dataset generalization.

Math solver ensemble. Yao et al. (2023) incorporated a problem typing subtask that combines the strengths of the tree-based solver and the LLM solver (ChatGLM-6B (Zeng et al., 2023)).

5 Analysis

5.1 LLMs's robustness in math

Patel et al. (2021) provided strong evidence that the pre-LLM MWP solvers, mostly LSTM-equipped encoder-decoder models, rely on shallow heuristics to achieve high performance on some simple benchmark datasets, then introduced a more challenging dataset, SVAMP, created by applying carefully chosen variations over examples sampled from preceding datasets. Stolfo et al. (2023) observed that, among non-instruction-tuned LLMs, the larger ones tend to be more sensitive to changes in the ground-truth result of a MWP, but not necessarily more robust. However, a different behavior exists in the instruction-tuned GPT-3 models, which show a remarkable improvement in both sensitivity and robustness, although the robustness reduces when problems get more complicated. Wei et al. (2023) assessed the robustness of several top-performing LLMs by augmenting the original problems in the curated CMATH dataset with distracting information. Their findings reveal that GPT-4 can maintain robustness while other models fail.

Zhou et al. (2023b) proposed a new dataset ROBUSTMATH to evaluate the robustness of LLMs in math-solving ability. Extensive experiments show that (i) Adversarial samples from higher-accuracy LLMs are also effective for attacking LLMs with lower accuracy; (ii) Complex MWPs (such as more solving steps, longer text, more numbers) are more vulnerable to attack; (iii) We can improve the robustness of LLMs by using adversarial samples in few-shot prompts.

5.2 Factors in influencing LLMs in math

The comprehensive evaluation conducted by Yuan et al. (2023) encompasses OpenAI's GPT series, including GPT-4, ChatGPT2, and GPT-3.5, along with various open-source LLMs. This analysis methodically examines the elements that impact the arithmetic skills of LLMs, covering aspects such as tokenization, pre-training, prompting techniques, interpolation and extrapolation, scaling laws, Chain of Thought (COT), and In-Context Learning (ICL).

Tokenization. This research underscores tokenization's critical role in LLMs' arithmetic performance (Yuan et al., 2023). Models like T5, lacking specialized tokenization for arithmetic, are less effective than those with advanced methods, such as Galactica (Taylor et al., 2022) and LLaMA, which show superior accuracy in arithmetic tasks. This indicates that token frequency in pre-training and the method of tokenization are key to arithmetic proficiency.

Pre-training Corpus. Enhanced arithmetic skills in LLMs correlate with the inclusion of code and LATEX in pre-training data (Yuan et al., 2023). Galactica, heavily utilizing LATEX, excels in arithmetic tasks, while models like Code-DaVinci-002, better at reasoning, lags in arithmetic, highlighting a distinction between arithmetic and reasoning skills.

Prompts. The nature of input prompts greatly affects LLMs' arithmetic performance (Liu et al., 2023a; Lou et al., 2023). Without prompts, performance drops (Yuan et al., 2023). Models like ChatGPT, which respond well to instructional system-level messages, demonstrate the importance of prompt type. Instruction tuning in pre-training also emerges as a significant factor (Yue et al., 2023).

Model Scale. There's a noted correlation between parameter count and arithmetic capability in LLMs (Yuan et al., 2023). Larger models generally perform better, but a performance plateau is observed, as shown by Galactica's similar outcomes at 30B and 120B parameters. However, this doesn't always mean superior performance, with smaller models like ChatGPT occasionally outperforming larger ones.

5.3 Perspectives of mathematics pedagogy

While machine learning emphasizes LLMs' problem-solving abilities in mathematics, in practical education, their primary role is to aid learning. Thus, the focus shifts from mere mathematical performance to a crucial consideration of LLMs' understanding of students' needs, capabilities, and learning methods.

Advantages of deploying LLMs in math education. Educators have observed the following benefits of leveraging LLMs for math education. (i) *LLMs foster critical thinking and problem-solving skills*, as they provide comprehensive solutions and promote rigorous error analysis (Matzakos et al., 2023); (ii) *Educators and students prefer LLM-generated hints* because of their detailed, sequential format and clear, coherent narratives (Gattupalli et al., 2023); (iii) *LLMs introduce a conversational style in problem-solving*, an invaluable asset in math education (Gattupalli et al., 2023); (iv) The impact of LLMs extends *beyond mere computational assistance*, offering deep insights and understanding *spanning diverse disciplines* like Algebra, Calculus, and Statistics (Rane, 2023).

Disadvantages of deploying LLMs in math education. (i) *Potential for misinterpretation.* Misinterpretation of students' queries or errors in providing explanations by LLMs could lead to confusion. Inaccurate responses might result in the reinforcement of misconceptions, impacting the quality of education (Yen and Hsu, 2023). (ii) *Limited understanding of individual learning styles.* LLMs may struggle to cater to diverse learning styles, as they primarily rely on algorithms and might not fully grasp the unique needs of each student. Some learners may benefit more from hands-on activities or visual aids that LLMs may not adequately address. Gattupalli et al. (2023) proposed that hints produced by GPT-4 could be excessively intricate for younger students who have shorter attention spans. (iii) *Privacy and data security issues*. Deploying LLMs involves collecting and analyzing substantial amounts of student data. Privacy concerns may arise if proper measures are not in place to safeguard this data from unauthorized access or misuse.

6 Challenges

Data-driven & limited generalization. The prevailing trend in current research revolves around the curation of extensive datasets. Despite this emphasis, there is a noticeable lack of robust generalization across various datasets, grade levels, and types of math problems. Examining how humans acquire math-solving skills suggests that machines may need to embrace continual learning to enhance their capabilities.

LLMs' brittleness in math reasoning. The fragility of LLMs in mathematical reasoning is evident across three dimensions. Firstly, when presented with questions expressed in varying textual forms (comprising words and numbers), LLMs exhibit inconsistent performance. Secondly, for identical questions, an LLM may yield different final answers through distinct reasoning paths during multiple trials. Lastly, pre-trained math-oriented LLMs are susceptible to attacks from adversarial inputs, highlighting their vulnerability in the face of manipulated data.

Human-oriented math interpretation. The current LLM-oriented math reasoning, such as chain-of-thoughts, does not take into account the needs and comprehension abilities of users, such as students. As an example, Yen and Hsu (2023) discovered that GPT-3.5 had a tendency to misinterpret students' questions in the conversation, resulting in a failure to deliver adaptive feedback. Additionally, research conducted by Gattupalli et al. (2023)

revealed that GPT-4 frequently overlooks the practical comprehension abilities of younger students. It tends to generate overly intricate hints that even confuse those students. Consequently, there is a pressing need for increased AI research that actively incorporates human factors into its design, ensuring future developments align more closely with the nuanced requirements of users.

7 Conclusion

This survey on LLMs for Mathematics delves into various aspects of LLMs in mathematical reasoning, including their capabilities and limitations. The paper discusses different types of math problems, datasets, and the persisting challenges in the domain. It highlights the advancements in LLMs, their application in educational settings, and the need for a human-centric approach in math education. We hope this paper will guide and inspire future research in the LLM community, fostering further advancements and practical applications in diverse mathematical contexts.

References

Chris Alvin, Sumit Gulwani, Rupak Majumdar, and Supratik Mukhopadhyay. 2017. Synthesis of solutions for shaded area geometry problems. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 14–19. AAAI Press.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of NAACL-HLT*, pages 2357–2367.

Jisu An, Junseok Lee, and Gahgene Gweon. 2023a. Does chatgpt comprehend the place value in numbers when solving math word problems? In *Proceedings of the Workshop "Towards the Future of AI-augmented Human Tutoring in Math Learning" co-located with The 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 3, 2023*, volume 3491 of *CEUR Workshop Proceedings*, pages 49–58.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023b. Learning from mistakes makes LLM better reasoner. *CoRR*, abs/2310.20689.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox. 2019. Holist: An environment for machine learning of higher-order theorem proving.

Yi Bin, Wenhao Shi, Yujuan Ding, Yang Yang, and See-Kiong Ng. 2023. Solving math word problems with reexamination. *CoRR*, abs/2310.09590.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of EMNLP*, pages 3313–3323.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2021a.

Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of ACL/IJCNLP*, volume ACL/IJCNLP 2021, pages 513–523.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021b. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of EMNLP*, pages 7889–7901.

Vincent Cheng and Yu Zhang. 2023. Analyzing ChatGPT's mathematical deficiencies: Insights and contributions. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 188–193.

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. *CoRR*, abs/2110.14168.

Aniruddha Deb, Neeva Oza, Sarthak Singla, Dinesh Khandelwal, Dinesh Garg, and Parag Singla. 2023. Fill in the blank: Exploring and enhancing LLM capabilities for backward reasoning in math word problems. *CoRR*, abs/2310.01991.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32):e2123433119.

Simon Frieder, Julius Berner, Philipp Petersen, and Thomas Lukasiewicz. 2023a. Large language models for mathematicians. *Internationale Mathematische Nachrichten*, 254:1–20.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023b. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Sai Gattupalli, William Lee, Danielle Allessio, Danielle Crabtree, Ivon Arroyo, Beverly Woolf, and Beverly Woolf. 2023. Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. In *Findings of ACL*, pages 5889–5903.

Sophia Gu. 2023. Llms as potential brainstorming partners for math and science problems. *CoRR*, abs/2310.10677.

Jesse Michael Han, Jason Rute, Yuhuai Wu, Edward W. Ayers, and Stanislas Polu. 2022. Proof artifact co-training for theorem proving with language models. In *Proceedings of ICLR*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*.

Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *CoRR*, abs/2304.09102.

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 *Proceedings of NeurIPS*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of EMNLP*, pages 523–533. ACL.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of ACL*, pages 37–42.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Trans. Assoc. Comput. Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of NAACL*, pages 1152–1157.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Zhenwen Liang, Dian Yu, Xiaoman Pan, Wenlin Yao, Qingkai Zeng, Xiangliang Zhang, and Dong Yu. 2023a. Mint: Boosting generalization in mathematical reasoning via multi-view fine-tuning. *CoRR*, abs/2307.07951.

Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kalyan. 2023b. Let GPT be a math tutor: Teaching math word problem solvers with customized exercise generation. *CoRR*, abs/2305.14386.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *CoRR*, abs/2305.20050.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of ACL*, pages 158–167.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, and Liang He. 2023b. Mathematical language models: A survey. *CoRR*, abs/2312.07622.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. 2023c. Improving large language model fine-tuning for solving math problems. *CoRR*, abs/2310.10047.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2023. Is prompt all you need? no. a comprehensive and broader view of instruction learning. *arXiv preprint arXiv:2303.10475*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023a. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*, abs/2310.02255.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of ACL/IJCNLP*, pages 6774–6786.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *Proceedings of ICLR*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023c. A survey of deep learning for mathematical reasoning. In *Proceedings of ACL*, pages 14605–14631.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *CoRR*, abs/2308.09583.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, pages 2263–2279.

Nikolaos Matzakos, Spyridon Doukakis, and Maria Moundridou. 2023. Learning mathematics with large language models: A comparative study with computer algebra systems and other tools. *International Journal of Emerging Technologies in Learning (iJET)*, 18(20):51–71.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of ACL*, pages 975–984.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of EMNLP*, pages 5807–5832.

Kole Norberg, Husni Almoubayyed, Stephen E. Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steven Ritter. 2023. Rewriting math word problems with large language models. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 7, 2023*, volume 3487 of *CEUR Workshop Proceedings*, pages 163–172.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus

Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of NAACL-HLT*, pages 2080–2094.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *Proceedings of ACL/IJCNLP*, pages 5870–5881.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Syed Rifat Raiyan, Md. Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2023. Math word problem solving by generating linguistic variants of problem statements. *CoRR*, abs/2306.13899.

Nitin Rane. 2023. Enhancing mathematical capabilities through chatgpt and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. *SSRN Electronic Journal*.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2023. Mathematical discoveries from program search with large language models. *Nature*, pages 1–3.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of EMNLP*, pages 1743–1752.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Mrinmaya Sachan, Avinava Dubey, and Eric P. Xing. 2017. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of EMNLP*, pages 773–784.

Mrinmaya Sachan and Eric P. Xing. 2017. Learning to solve geometry problems from natural language demonstrations in textbooks. In *Proceedings of *SEM @ACM*, pages 251–261.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. ARB: advanced reasoning benchmark for large language models. *CoRR*, abs/2307.13692.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of EMNLP*, pages 1466–1476.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An independent evaluation of chatgpt on mathematical word problems (MWP). In *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023*, volume 3433 of *CEUR Workshop Proceedings*.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of ACL*, pages 545–561.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Alberto Testolin. 2023. Can neural networks do arithmetic? A survey on the elementary numerical skills of state-of-the-art deep learning models. *CoRR*, abs/2303.07735.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*.

Shyam Upadhyay and Ming-Wei Chang. 2017. Annotating derivations: A new evaluation strategy and dataset for algebra word problems. In *Proceedings of EACL*, pages 494–504.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

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 *Proceedings of ICLR*.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of EMNLP*, pages 845–854.

Zichao Wang, Andrew S. Lan, and Richard G. Baraniuk. 2021. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of EMNLP*, pages 5986–5999.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. CMATH: can your language model pass chinese elementary school math test? *CoRR*, abs/2306.16636.

Makarius Wenzel, Lawrence C Paulson, and Tobias Nipkow. 2008. The isabelle framework. In *Theorem Proving in Higher Order Logics: 21st International Conference, TPHOLs 2008, Montreal, Canada, August 18-21, 2008. Proceedings 21*, pages 33–38. Springer.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023. An empirical study on challenging math problem solving with GPT-4. *CoRR*, abs/2306.01337.

Ryutaro Yamauchi, Sho Sonoda, Akiyoshi Sannai, and Wataru Kumagai. 2023. LPML: llm-prompting markup language for mathematical reasoning. *CoRR*, abs/2309.13078.

Kaiyu Yang and Jia Deng. 2019. Learning to prove theorems via interacting with proof assistants.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. 2023. GPT can solve mathematical problems without a calculator. *CoRR*, abs/2309.03241.

Jie Yao, Zihao Zhou, and Qiufeng Wang. 2023. Solving math word problem with problem type classification. In *Proceedings of NLPCC*, volume 14304, pages 123–134.

An-Zi Yen and Wei-Ling Hsu. 2023. Three questions concerning the use of large language models to facilitate mathematics learning. *CoRR*, abs/2310.13615.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *CoRR*, abs/2309.12284.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *CoRR*, abs/2304.02015.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *CoRR*, abs/2309.05653.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *Proceedings of ICLR*.

Beichen Zhang, Kun Zhou, Xilin Wei, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2023a. Evaluating and improving tool-augmented computation-intensive math reasoning. *arXiv preprint arXiv:2306.02408*.

Mengxue Zhang, Zichao Wang, Zhichao Yang, Weiqi Feng, and Andrew S. Lan. 2023b. Interpretable math word problem solution generation via step-by-step planning. In *Proceedings of ACL*, pages 6858–6877.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. Minif2f: a cross-system benchmark for formal olympiad-level mathematics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023a. Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. *CoRR*, abs/2308.07921.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2023b. Mathattack: Attacking large language models towards math solving ability. *CoRR*, abs/2309.01686.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2023. Solving math word problems via cooperative reasoning induced language models. In *Proceedings of ACL*, pages 4471–4485.

Mingyu Zong and Bhaskar Krishnamachari. 2023. Solving math word problems concerning systems of equations with GPT-3. In *Proceedings of AAAI*, pages 15972–15979.