An Analysis of Large Language Models and LangChain in Mathematics Education

Conference Paper · October 2023

DOI: 10.1145/3633598.3633614

2 authors:

Fatih Soygazi
Aydın Adnan Menderes University
39 PUBLICATIONS **88** CITATIONS
SEE PROFILE

Damla Oguz
Izmir Institute of Technology
13 PUBLICATIONS **122** CITATIONS
SEE PROFILE

An Analysis of Large Language Models and LangChain in Mathematics Education

Fatih Soygazi*
fatih.soygazi@adu.edu.tr
Department of Computer Engineering, Aydın Adnan Menderes University
Aydın, Turkey

Damla Oguz*
damlaoguz@iyte.edu.tr
Department of Computer Engineering, İzmir Institute of Technology
İzmir, Turkey

ABSTRACT

The development of large language models (LLMs) has led to the consideration of new approaches, particularly in education. Word problems, especially in subjects like mathematics, and the need to solve these problems by collectively addressing specific stages of reasoning, have raised the question of whether LLMs can be successful in this area as well. In our study, we conducted analyses by asking mathematics questions especially related to word problems using ChatGPT, which is based on the latest language models like Generative Pretrained Transformer (GPT). Additionally, we compared the correct and incorrect answers by posing the same questions to LLMMathChain, a mathematics-specific LLM based on the latest language models like LangChain. It was observed that the answers obtained were more successful with ChatGPT (GPT 3.5), particularly in the field of mathematics. However, both language models were found to be below expectations, particularly in word problems, and suggestions for improvement were provided.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

ChatGPT, LangChain, Large Language Models (LLMs), Mathematics Education

ACM Reference Format:
Fatih Soygazi and Damla Oguz. 2023. An Analysis of Large Language Models and LangChain in Mathematics Education. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

LLMs have brought about a significant transformation in natural language processing (NLP) by enabling machines the ability to understand human language in an unprecedented manner. These models can examine vast amounts of text data and learn language patterns, thereby generating human-like responses to questions. The release of Open AI's GPT-3 [3] and ChatGPT has contributed to the widespread popularity of large language LLMs. While it may not be feasible to train an LLM from scratch due to limited computational resources, individuals can still harness the power of pretrained LLMs to develop exciting and innovative projects.

Chatbots using LLMs can be customized to serve diverse objectives, including education. The ability to obtain accurate responses by utilizing appropriate prompts through prompt engineering demonstrates that chatbots can be utilized as educational assistants. Therefore, advanced chatbot models such as ChatGPT are commonly employed for this purpose. However, it is important to note that ChatGPT does not consistently provide completely accurate answers to all math questions. According to Shakarian et al.[16], ChatGPT's performance falls notably below the 60% accuracy achieved by state-of-the-art algorithms designed specifically for solving math word problems. The disparity arises due to the inherent differences between AI systems and human problem-solving abilities. AI models like ChatGPT require sufficient time to develop a comprehensive understanding of complex math problems and reach a level of proficiency comparable to that of a human expert. Because the human-level question answering requires vast amount of math questions and the related answers that could be fed into GPT by time.

During this developmental stage, an alternative NLP framework called LangChain[1] has been introduced, which also focuses on Language Model (LM) applications. The concept behind LangChain is to rapidly build diverse applications for various domains by leveraging LMs. LangChain is data-aware, meaning it can connect an LM to other data sources. It incorporates abstractions that work alongside LMs, offering a collection of implementations for each abstraction.

Applying the principles of LangChain to math, we can utilize commonly available LMs by feeding them math datasets, potentially yielding improved results. Consequently, it becomes plausible to link the Hugging Face and OpenAI endpoints to LangChain, which takes into account math datasets. This way, the results achieved with ChatGPT can be effortlessly obtained with LangChain by employing a specific language model such as LLMMathChain[2].

The objective of this paper is to address math word problems and math numerical computational problems using ChatGPT and LangChain, with a focus on analyzing the performance of these Language Models (LLMs). Numerical computational problems typically

*Both authors contributed equally to this research.

Conference acronym 'XX, June 13–15, 2023, İstanbul, Turkey

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1]https://github.com/langchain-ai/langchain
[2]https://python.langchain.com/docs/modules/chains/additional/llm_math

have established solution methods, whereas word problems require accurate comprehension before they can be solved. Consequently, the accuracy of solutions may vary, and this variation is thoroughly examined by considering different question and error types within the context of math problems.

The structure of this paper is as follows: Section 2 presents the related work, Section 3 covers the analysis of prompts, including the categorizations of mathematics problems and error types. Section 4 provides a discussion of the study. Lastly, Section 5 concludes the paper by presenting the key findings, summarizing the conclusions, discussing the future research.

2 RELATED WORK

In the field of mathematics, there has been a significant focus on working with current Large Language Models (LLMs) during the process of solving word and numerical computational problems. Particularly in the case of word problems, it is crucial to convert expressions that humans can understand into a logic that can be comprehended by machines. Zhang et al. [27] indicate that existing systems in the literature have shown promising results when tested on self-crafted and small-scale datasets. However, their performance declines significantly when applied to large and diverse datasets, indicating the need for significant improvement in current mathematics word problem (MWP) solvers. Zhang et al. [27] present a comprehensive survey that aims to provide a clear and comprehensive overview of automatic mathematics problem solvers. The survey primarily focuses on algebraic word problems, highlighting the extracted features and proposed techniques to bridge the semantic gap. The performance of these solvers is compared using publicly accessible datasets. Additionally, the survey explores automatic solvers for other types of mathematics problems, including geometric problems that involve diagram comprehension. The technological trend in solving MWPs have been passed through rule-based matching from 1960 to 2010, semantic parsing and statistical learning from 2011 to 2017 and deep-reinforcement learning from 2017 to nowadays. The rule-based solvers [2, 6, 14, 17, 26] heavily depend on human interventions and are restricted to addressing a predetermined set of scenarios. Hence, the rewritten questions or addition of a simple subproblem to the problem may cause the inability to solve the problem properly. During the second stage of development, MWP solvers started incorporating semantic parsing techniques. The aim was to convert sentences from problem statements into structured logic representations, enabling easier quantitative reasoning. These methods employ diverse strategies such as feature engineering and statistical learning to enhance performance. These methods have focused on publicly available or manually collected datasets. The language understanding of the questions has been improved with respect to rule-based solvers but still the understanding might not directly enable to solve the problems in an iterative manner of subproblems. Around 2017, the impact of semantic parsing by new approaches for inferencing [9] and the increase in deep learning [22] have been the workforce of MWP solvers.

Recurrent Neural Network (RNN) [22], encoder-decoder network [4], deep Q network [21], reinforcement learning [8] based approaches are the preliminary approaches of deep learning until the rise of Transformer networks [18]. MWP-BERT and MWP-RoBERTa [11], SciBERT and MathBERT [13] enable BERT based approaches for mathematics word problems.

Another MWP approach is the tree-based approaches [12, 20, 25] and combination of tree-based approaches with Transformers [15]. Tree-based approaches are commonly used to represent arithmetic expressions as binary tree structures. These approaches aim to construct an equivalent tree structure in a bottom-up manner, gradually transforming the derivation of the arithmetic expression. Notably, tree-based approaches do not require additional annotations like equation templates, tags, or logic forms. The algorithmic framework of these approaches involves two stages: the extraction of quantities from the text to form the bottom level of the tree and the enumeration of syntactically valid candidate trees with different structures and internal nodes. In the second stage, a scoring function is used to select the best matching candidate tree, which determines the final solution.

The state-of-the-art Transformer model, GPT, has been considered for mathematics problems recently. Liang et al.[10] introduce a novel approach to distill mathematics word problem-solving capabilities from large language models (LLMs) into smaller, more efficient student models for leveraging LLMs. The method involves generating an exercise book to evaluate student models and providing customized training exercises based on the specific learning needs of each student model. Experimental results demonstrate that this approach achieves higher accuracy than LLMs (such as GPT-3 and PaLM [5]) across three different benchmarks while utilizing significantly fewer parameters.

When LangChain is a newly proposed technology, there is not a paper that compares ChatGPT and LangChain for mathematics problems in detail to the best of our knowledge. In this paper, we propose a categorization of mathematics question types and the corresponding errors, and use prompts to analyze the comparison of ChatGPT (GPT-3.5) and LangChain.

3 METHODOLOGY

Prompt analysis is the main methodology of this study which refers to the process of examining and understanding a given prompt or question in various contexts, often in the context of natural language processing (NLP) tasks or AI models. Since the context is mathematics in this study, we present a prompt analysis of ChatGPT and LangChain by using different mathematical questions.

In this section, we begin by introducing and categorizing different question types, as well as addressing the errors that occurred in these questions. Subsequently, we present the questions used in the prompt analysis along with their results.

3.1 Categorization of Question Types and Error Types

We posed two types of questions to ChatGPT and LangChain: word problems and numerical computational problems. Word problems are typically considered as textual problems that can be resolved by utilizing mathematical concepts, rules, or techniques [19]. We use numerical computational problems to describe problems that specifically require performing numerical calculations or computations.

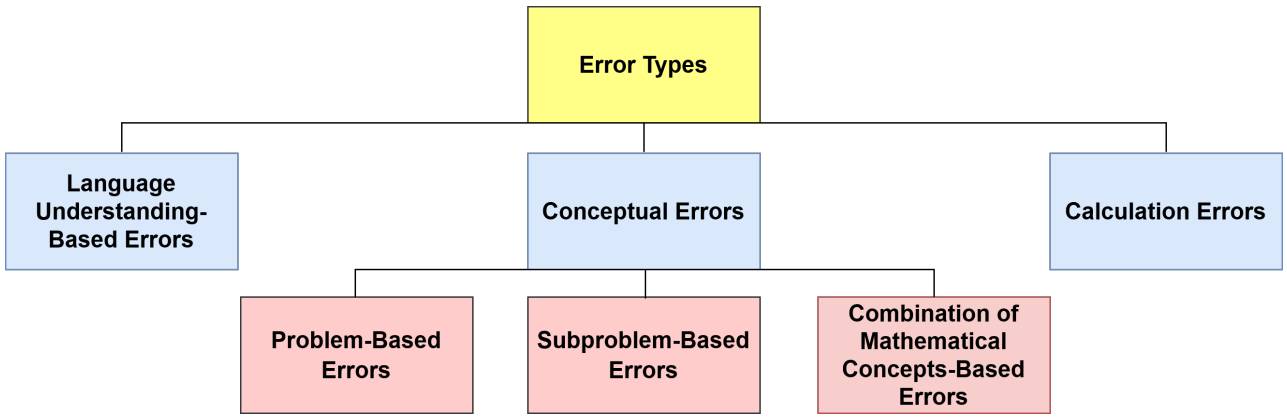

Figure 1: Categorization of Error Types.

We categorize the error types as language understanding-based errors, conceptual errors, and calculation errors. Language understanding-based errors in this paper refer to errors caused by a lack of understanding of the problem. These errors are resolved by rewriting or rephrasing the problem.

Conceptual errors are used to describe the errors related with applying the wrong theorem or formula in problem solving [7, 23]. We also categorize conceptual errors into three sub-categories: problem-based, subproblem-based and combination of mathematical concepts-based errors. Problem-based conceptual errors indicate a mistake made in applying the theorem or formula to the solution as a whole. It can also be caused by a lack of knowledge of ChatGPT or LLMMathChain. Subproblem-based conceptual errors refer to errors made specifically in the solution of sub-parts of the problem. Combination of mathematical concepts-based errors occur due to a lack of correct utilization of a combination of different mathematical topics. Calculation errors involve mistakes in performing arithmetic operations. Our categorization of error types is shown in Figure 1.

3.2 Prompt Analysis

In this subsection, we introduce the questions used in our prompt analysis, along with their corresponding outcomes. Figure 2 provides an overview of the questions employed in the prompt analysis.

Question 1, which is a word problem, was answered correctly by ChatGPT as 6, while it was answered incorrectly by LLMMathChain. LLMMathChain made the following evaluation: 5 - 2 + 4, which was thought to us the indication a language understanding problem. We rewrote the question as Question 1' to support our idea. In that case, LLMMathChain understood the question and responded correctly as ChatGPT did.

Question 2 is also a word problem with the correct answer of 14. ChatGPT provided answers of 8, -8, 14, and 8, respectively. Although ChatGPT correctly constructed the equation, it made a calculation error, resulting in incorrect answers in some attempts. On the other hand, LLMMathChain had a syntax error, which can be attributed to a language understanding-based error. We think that there might be an issue somewhere when using the models employed in LLMMathChain.

The type of Question 3 is a word problem, and the correct answer is 66. Initially, ChatGPT provided an incorrect answer because it couldn't formulate the equation of the solution accurately, resulting in a problem-based error. When we asked the question again, ChatGPT provided another incorrect answer. In the third attempt, ChatGPT was able to construct an equation that was close to the correct one, with a result of 65. However, the answer still remained incorrect. Similarly, LLMMathChain also provided incorrect answers due to a conceptual error related to the problem itself. The difference between ChatGPT and LLMMathChain was that LLMMathChain consistently gave the same incorrect answer.

Question 4 is a word problem with the correct answer of 7. ChatGPT correctly answered this question. However, LLMMathChain performed the following evaluation: $(4*(20-3))/2+(20/4)**(1/2)$ which resulted in an incorrect answer due to a subproblem-based error, which is a subcategory of conceptual error. Question 5 is yet another word problem with subproblems, and the correct answer is 82. Although ChatGPT gave the correct answer, LLMMathChain did the following equation (52/2+52/3+2)+(40+(52/2+52/3+2)) and responded to the question wrongly. This issue is similar to Question 4, where LLMMathChain's incorrect answer was also attributed to an error in managing the subproblems.

Question 6 is a numerical computational problem which covers two different topics of mathematics, and a method needs to be applied while solving this question. The two different topics should be considered together. First, it is necessary to determine whether the content inside the absolute value is positive or negative. After this determination, the part inside the absolute value should come out as a function, not numerically. However, ChatGPT calculates the content inside the absolute value numerically first, which leads to an error. We define this type of error as errors based on the combination of mathematical concepts. The same situation exists here as well in ChatGPT. In short, Question 6 was answered incorrectly by both ChatGPT and LLMMathChain because they do not know how to combine different topics in a problem solution. However, this knowledge can be taught.

Question 7 is another numerical computational problem with the correct answer of -11. It involves different mathematical topics, namely absolute value, and derivative. In this case, ChatGPT

Question 1	Sue has 5 apples and Jennifer has 4 apples. Sue gives 2 apples to Jennifer. How many apples will Jennifer have?
Question 1′	Sue has 5 apples and Jennifer has 4 apples. Sue gives 2 apples to Jennifer. How many apples will Sue have?
Question 2	Jennifer can complete the whole task in 3x days, and Ben can complete the whole task in x/2 days. If Jennifer and Ben can together complete the whole task in 12 days, how many days would it take for Ben to complete the same task alone?
Question 3	In a market, when you purchase 5 toys, you receive an additional 2 free toys. A person who has bought a total of 92 toys from the market, how many of these toys did they buy by paying for them?
Question 4	Sue and April, two sisters, have a total of 20 apples. The number of Sue's apples is four times that of April's. If Sue gives Aria half of her apples minus three, and April gives Aria the square root of her apples, how many apples will Aria have?
Question 5	The total number of students in a classroom is 52. The students sit in pairs on 5 desks and in groups of three on the other desks. The number of desks in the classroom plus two is equal to Alia's age. Amy is 40 years older than Alia. What is the sum of Amy and Alia's ages?
Question 6	If $y = \|4x - 5\| + x^2$, for $x = -1$, $\frac{dy}{dx} = ?$
Question 7	If $f(x) = \|x^3 + x^2 + 3x - 1\|$ then $f'(-2) = ?$

Figure 2: Questions in the Prompt Analysis.

provided the correct answer. However, LLMMathChain encountered the same issue as in Question 6 and answered the question incorrectly due to errors based on the combination of mathematical concepts.

Table 1 provides a summary of the questions that we asked, and the results obtained from ChatGPT and LLMMathChain. ChatGPT was more successful than LLMMathChain in these questions. Table 2 displays the match between question types and error types. As can be seen from the table, ChatGPT did not exhibit any language-understanding-based or subproblem-based errors in our questions. However, LLMMathChain experienced errors of all types except for calculation errors.

4 DISCUSSION

In mathematics, the answers are typically deterministic, meaning there's a specific correct answer. Hence, the mathematics softwares must provide consistent and accurate responses to prompts. However, since LMs are statistical in nature, we can change the methods we use to reach a solution. These changes can cause the answers to vary each time, which goes against the nature of mathematics. Similarly, ChatGPT, by design, does not always give the same answer to the same question. It's based on a Transformer model which is designed to generate diverse and creative outputs. This is particularly useful for tasks such as writing a story or generating a conversation, where generative AI is the primary consideration. When it comes to mathematics questions, the model doesn't solve them in the same manner as a calculator. Instead, it generates responses based on patterns it learned during training. So, while it often provides correct answers to mathematical questions, it might not always be consistent because it does not "understand" mathematics in the way humans do. Its responses are based on patterns rather than performing calculations.

Figure 3 represents the combination of Natural Language Understanding (NLU) and Natural Language Generation (NLG) in an LLM. During the NLU stage, expressions written as word problems or numerical computational problems are transformed into a form that can be solved using mathematical methods. NLU is responsible for converting the problem into symbolic notation. Afterwards, NLG, with the help of extensive training data, should be able to apply the solutions it has learned to the problems expressed in the NLU stage. To overcome the problem of non-deterministic features of LLMs like ChatGPT, it requires training LLMs on diverse mathematical datasets spanning a large number of questions from different mathematics topics.

Reinforcement learning (RL) is a branch of machine learning concerned with training agents to make a series of decisions within an environment. Its goal is to maximize the overall reward they receive. It draws inspiration from the way humans and animals learn by engaging in trial-and-error interactions with their surroundings. ChatGPT benefits from reinforcement learning in its reward model called Reinforcement Learning from Human Feedback (RLHF) [24]. ChatGPT's initial training is done using supervised fine-tuning, where human AI trainers provide conversations and model responses via RLHF. The advantage of this approach is to teach the LLM with the situations where it has not encountered until that time. The AI trainer ranks different model-generated responses based on their quality and may change the knowledge or

Table 1: Question Types and the Result Comparison

Questions	Question Type	ChatGPT	LLMMathChain
Question 1	Word Problem	Correct	Wrong
Question 1´	Word Problem	Correct	Correct
Question 2	Word Problem	Wrong	Wrong
Question 3	Word Problem	Wrong	Wrong
Question 4	Word Problem	Correct	Wrong
Question 5	Word Problem	Correct	Wrong
Question 6	Numerical Comp. P.	Wrong	Wrong
Question 7	Numerical Comp. P.	Correct	Wrong

Table 2: Question-error Type Match

Error Types	ChatGPT	LLMMathChain
Language Understanding-Based Error	-	Questions 1 & 2
Problem-Based Error	Question 3	Question 3
Subproblem-Based Error	-	Questions 4 & 5
Calculation Error	Question 2	-
Combination of Math. Concepts-Based Error	Question 6	Questions 6 & 7

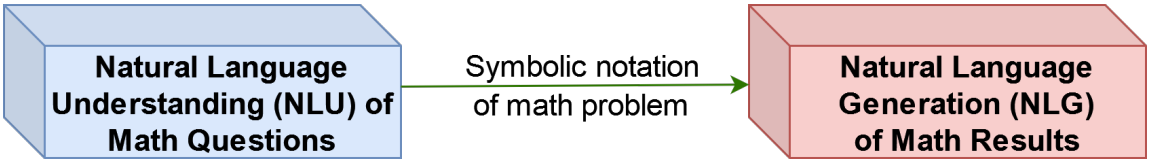

Figure 3: Current LLM Approach for Mathematics Problems.

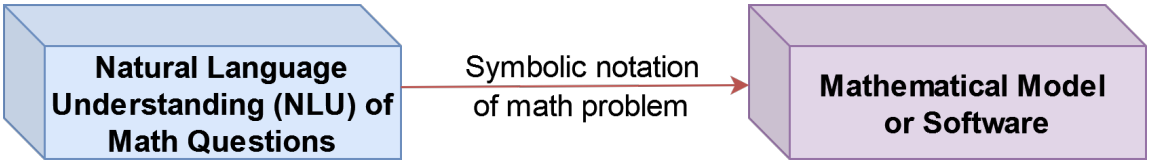

Figure 4: Current LangChain Approach for Mathematics Problems.

beliefs. These rankings serve as reward signals to LLM for optimizing the expected output. Collective expertise of humans generally helps ChatGPT to gain experience positively. But scientific areas like mathematics is not the case [1] because mathematics is not a social area where beliefs or truths must change the facts. It is a scientific area that techniques of solving a question is algorithmic and these algorithms should be learnt by some amount of similar question-solution pairs. This is the way a student works for his exams to solve lots of questions with varieties related to a specific topic.

Figure 4 summarizes how LLMMathChain operates to solve mathematical problems. When compared to Figure 3, LangChain aims to achieve more successful results by using mathematical software or a model specifically trained for mathematics. Hence, the disadvantage of LLMs can be mitigated by leveraging the deterministic nature of mathematics. LangChain allows you to use different LLMs served by OpenAI or Hugging Face endpoints with various datasets or softwares in combination to improve the performance of the system. For example, LLMMathChain could be used with LLM named as text-davinci that is used by ChatGPT-3.5. Hence, the question answering capabilities of ChatGPT like syntactic or semantical inferences could be served while the additional mathematical knowledge could be infused to a mathematics specific software or model.

There are some efforts to link LLMs with actual mathematics softwares like GPT and Wolfram Alpha3 [3]. This is a demonstration of a conversational agent using OpenAI GPT-3.5 and LangChain powered with WolframAlpha. The addition of community accepted softwares usage with ChatGPT will help the solution of mathematics problems in a more accurate and deterministic manner.

Mathematics is often thought of as the language of logic. Mathematics provides a precise and formal system of symbols, notations

[3]https://huggingface.co/spaces/JavaFXpert/Chat-GPT-LangChain

and rules that enable logical thinking and precise communication. It allows us to express and manipulate abstract concepts, formulate logical arguments, and draw conclusions based on logical deductions. While ML is an inductive reasoning-based approach, it is the difficulty of LLMs to handle the problems in a deductive manner. As logical systems such as propositional logic or first-order logic have a way of logical consequences for solution, these consequences must also be learnt by LLMs with a vast amount of similar deduction-based problems. LangChain is expected to be capable of solving these problems by time in combination with GPT.

5 CONCLUSION

In this study, an analysis and comparison of solving different mathematics problems, particularly word problems and numerical computational problems, have been presented using state-of-the-art LLMs and LangChain. The process of learning mathematics with LLMs differs significantly from the nature of mathematics itself. Moreover, the generative and statistical nature of LLMs is not inherently well-suited for deterministic fields like mathematics. However, it is projected that with novel approaches such as LangChain, the accuracy of mathematical question answering can be improved over time. Therefore, it is evident that by nourishing the success of LLMs in language comprehension aligned with the nature of mathematics, a successful mathematics assistant can be developed.

In future studies, all mathematics models and softwares that can be adapted to LangChain will be utilized with current LLMs, and the effectiveness of the Chain of Thought (CoT) method for unanswered questions will be determined. Additionally, considering the utilization of tree-based methods, particularly decision trees for explainability in machine learning, efforts will be made to investigate the explainability of the mathematics question answering method within the mathematics chain.

REFERENCES

[1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? arXiv:2303.12767 [cs.CL]

[2] Daniel Bobrow. 1964. Natural language input for a computer problem solving system. (1964).

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[4] Ting-Rui Chiang and Yun-Nung Chen. 2018. Semantically-aligned equation generation for solving and reasoning math word problems. *arXiv preprint arXiv:1811.00720* (2018).

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[6] Charles R Fletcher. 1985. Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, & Computers* 17, 5 (1985), 565–571.

[7] James Hiebert. 2013. *Conceptual and procedural knowledge: The case of mathematics*. Routledge.

[8] Danqing Huang, Jing Liu, Chin-Yew Lin, and Jian Yin. 2018. Neural math word problem solver with reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*. 213–223.

[9] Danqing Huang, Shuming Shi, Chin-Yew Lin, and Jian Yin. 2017. Learning Fine-Grained Expressions to Solve Math Word Problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 805–814. https://doi.org/10.18653/v1/D17-1084

[10] Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kalyan. 2023. Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation. *CoRR* abs/2305.14386 (2023). https://doi.org/10.48550/arXiv.2305.14386 arXiv:2305.14386

[11] Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. MWP-BERT: Numeracy-Augmented Pre-training for Math Word Problem Solving. arXiv:2107.13435 [cs.AI]

[12] Qianying Liu, Wenyv Guan, Sujian Li, and Daisuke Kawahara. 2019. Tree-structured Decoding for Solving Math Word Problems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2370–2379. https://doi.org/10.18653/v1/D19-1241

[13] Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023. A Symbolic Framework for Systematic Evaluation of Mathematical Reasoning with Transformers. arXiv:2305.12563 [cs.CL]

[14] Anirban Mukherjee and Utpal Garain. 2008. A review of methods for automatic understanding of natural language mathematical problems. *Artificial Intelligence Review* 29 (2008), 93–122.

[15] Alexander Scarlatos and Andrew Lan. 2023. Tree-Based Representation and Generation of Natural and Mathematical Language. arXiv:2302.07974 [cs.CL]

[16] Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). arXiv:2302.13814 [cs.CL]

[17] James R Slagle. 1965. Experiments with a deductive question-answering program. *Commun. ACM* 8, 12 (1965), 792–798.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Lieven Verschaffel, Stanislaw Schukajlow, Jon Star, and Wim Van Dooren. 2020. Word problems in mathematics education: A survey. *ZDM* 52 (2020), 1–16.

[20] Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. 2018. Translating a Math Word Problem to a Expression Tree. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1064–1069. https://doi.org/10.18653/v1/D18-1132

[21] Lei Wang, Dongxiang Zhang, Lianli Gao, Jingkuan Song, Long Guo, and Heng Tao Shen. 2018. Mathdqn: Solving arithmetic word problems via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[22] Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep Neural Solver for Math Word Problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 845–854. https://doi.org/10.18653/v1/D17-1088

[23] Widodo Winarso and Toheri Toheri. 2021. An analysis of students' error in learning mathematical problem solving: The perspective of David Kolb's theory. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, 1 (2021), 139–150.

[24] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

[25] Zhipeng Xie and Shichao Sun. 2019. A Goal-Driven Tree-Structured Neural Model for Math Word Problems. In *Ijcai*. 5299–5305.

[26] Bakman Yefim. 2007. Robust Understanding of Word Problems with Extraneous Information. arXiv:math/0701393 [math.GM]

[27] Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai, and Heng Tao Shen. 2020. The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2287–2305. https://doi.org/10.1109/TPAMI.2019.2914054