The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Analyzing ChatGPT's Mathematical Deficiencies: Insights and Contributions

Vincent Cheng
Morrison Academy
vincentcheng236@gmail.com

Zhang Yu
National Central University
phoenix000.taipei@gmail.com

Abstract

In this study, we assess ChatGPT, OpenAI's latest conversational chatbot and large language model (LLM), on its performance in elementary-grade arithmetic and logic problems. Despite its impressive coherence in natural language processing and ability to follow instructions, our findings indicate that ChatGPT still has room for improvement in mathematical tasks. To evaluate its performance, we used six math and logic datasets, including SingleEq, AddSub, SVAMP, MultiArith, Simple Arithmetic and counting, and Arithmetic (word variation), and found that ChatGPT performed better than previous models such as InstructGPT and Minerva. However, our arithmetic dataset, which includes two- to seven-digit equations, revealed that ChatGPT's accuracy in solving addition problems decreased from 100% to 64%, with simple arithmetic errors such as not carrying over in addition being a common issue. Additionally, the model struggled with basic multi-step word problems. To address this, we propose a novel benchmark for evaluating LLMs' mathematical abilities. Further research is needed for LLMs to reach the level of mathematical reasoning comparable to their natural language processing abilities. Overall, our study highlights the need for continued improvement in LLMs' mathematical abilities to make them more effective in real-world applications.

Keywords: Large language models, reasoning capabilities

1 Introduction

Pretrained language models (PLMs) have revolutionized natural language processing, achieving impressive performance on various tasks, from sentiment analysis to question answering and text generation. With the development of large language models (LLMs), the capabilities of PLMs have grown even further, with models such as GPT-3 boasting over 100 billion parameters [Brown et al., 2020]. ChatGPT, a conversational chatbot and LLM developed by OpenAI, has become one of the most popular language models, with over 100 million users in under three months. However, while these models excel in language processing, they may lack the ability to reason mathematically and logically, as observed in previous models such as BART [Patel et al., 2021, Wang et al., 2021, Roy and Roth, 2016].

In this paper, we present a study of the mathematical and logical capabilities of ChatGPT, focusing on simple arithmetic, elementary-grade level math word problems, and logic problems. While previous research has analyzed ChatGPT's performance on advanced math problems with proofs from college-level pure math courses [Frieder et al., 2023], our research is unique as it presents a detailed analysis of ChatGPT's performance on simple mathematical and logical reasoning tasks. Our study evaluates ChatGPT's mathematical reasoning abilities, which have not been analyzed in previous research.

Moreover, while Borji [2023] briefly touches on various topics such as mathematical reasoning, hallucination, and bias, our analysis focuses solely on the model's ability to reason mathematically and logically. Our research aims to specifically contribute to the evaluation of LLMs' mathematical and logical capabilities.

Our study makes several contributions to the evaluation of ChatGPT's mathematical and logical capabilities:

1. We conducted a comprehensive assessment of ChatGPT's ability to reason mathematically and logically on simple tasks, comparing its performance with other LLMs of comparable parameter sizes.

2. We designed a word variation experiment to investigate ChatGPT's computational ability,

showing that the model's performance may depend on specific patterns in the pre-training corpus and that it has limitations in generalizing more common computational rules.

3. We evaluated ChatGPT's performance using both the commonly used Accuracy metric and the Average Percent Error (APE) metric, revealing that ChatGPT has the capability of estimation, even if it is not always accurate in some computational tasks.

4. We conducted an error analysis of ChatGPT's performance on some mathematical tasks, identifying "adding one extra digit" as a common type of error that deserves further investigation.

2 Methods

2.1 Datasets

We evaluated ChatGPT's performance on existing datasets from previous studies, which include:

1. SingleEq [Koncel-Kedziorski et al., 2015]
2. AddSub [Hosseini et al., 2014]
3. SVAMP [Patel et al., 2021]
4. MultiArith [Roy and Roth, 2016]

These datasets consist of simple single-step arithmetic problems written in word problem format or requiring multiple arithmetic steps to solve. Additionally, we extended the arithmetic and counting experiments from Wang et al. [2021] to include addition, subtraction, and multiplication problems with two to seven digits and evaluated multiple ranges for counting. We also created a Word Variation dataset by modifying the arithmetic problems and replacing the original Arabic numbers with English words, as detailed in section 4.2. Our datasets for arithmetic are created using a random number generator and word variations are generated using the num2words library from Python. They will be released to the public in the future.

2.2 Metrics

We used two metrics to evaluate ChatGPT's performance on these datasets: Accuracy and Average Percent Error (APE). The percent error for each sample is calculated using the following formula:

$$\text{Percent error} = \frac{|\text{Response} - \text{Actual answer}|}{\text{Actual answer}}$$

2.3 Experimental Setup

We conducted our experiments on the January 30th version of ChatGPT, using PyChatGPT [terry3041, 2023] to automate its use. For each sample, we prompted ChatGPT with the instruction, "Respond with only the answer to the following question: ..." and discarded any responses that were noisy or contained more than just the answer.

An example prompt, question, and response from ChatGPT are shown in Figure 1.

Due to the usage limits of ChatGPT, we were only able to use 100 test cases for each dataset during evaluation. However, on some tasks where ChatGPT performed poorly, we conducted at least three experiments and took the median of the results.

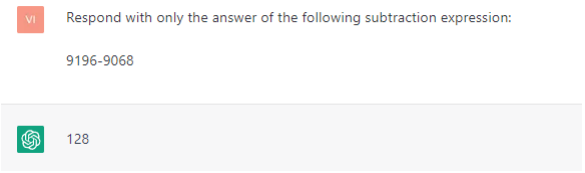

Figure 1: Example prompt and response

3 Results and discussion

3.1 Word Problem Performance Comparison

Table 2 summarizes the performance of ChatGPT on various word problem datasets, including AddSub, SingleEq, SVAMP, and MultiArith, as well as the performance of InstructGPT [Ouyang et al., 2022] and PaLM [Chowdhery et al., 2022] on the same datasets. The results of InstructGPT are taken from Kojima et al. [2022], while the results of PaLM are taken from Zhou et al. [2022].

ChatGPT performs relatively well on single-step word problems from the AddSub and SingleEq datasets. However, the SVAMP and MultiArith datasets have increased problem complexity, requiring more arithmetic operations than the first two datasets, and ChatGPT's performance decreases significantly on these tasks. Specifically, ChatGPT only achieves an accuracy of 64% on the SVAMP dataset.

We find that the problems in SVAMP require a higher level of comprehension compared to the other datasets, which are more straightforward. For example, the question "The grasshopper, the frog, and the mouse had a jumping contest. The grasshopper jumped 9 inches. The mouse jumped

Dataset	Prompt	Answer
SingleEq	The sum of three consecutive odd numbers is 69. What is the smallest of the three numbers?	21
AddSub	Joan found 70 seashells on the beach. She gave Sam some of her seashells. She has 27 seashells. How many seashells did she give to Sam?	43
SVAMP	Tiffany was collecting cans for recycling. On Monday she had 7 bags of cans. The next day she found 12 more bags worth of cans. How many more bags did she find on the next day than she had on Monday?	5
MultiArith	Kaleb was collecting cans for recycling. On Saturday he filled 5 bags up and on Sunday he filled 5 more bags. If each bag had 4 cans in it, how many cans did he pick up total?	40
Arithmetic	7342+3492	10834
Counting	How many ”i”s are there in the following string: ”i i i i i i i i i i”?	11
Arithmetic (word variation)	seven thousand, three hundred and forty-two plus three thousand, four hundred and ninety-two	10834

Table 1: Examples from each dataset

Model Name	Accuracy(%)			
	AddSub	SingleEq	SVAMP	MultiArith
InstructGPT	74.7	78.7	63.7	79.3
Minerva (PaLM)	91.9	-	-	**94.7**
ChatGPT	**94.0**	**89.0**	**64.0**	84.0

Table 2: Accuracy of ChatGPT and previous models on word problem datasets

3 inches lesser than the frog who jumped 33 inches farther than the grasshopper. How far did the mouse jump?” requires keeping track of the position of all three animals given their relative positions. ChatGPT answered this incorrectly with ”15” while the correct answer was ”39”.

It is worth noting that ChatGPT’s performance outperforms InstructGPT on most tasks, even without the chain-of-thought prompting used to elicit multi-step reasoning. These results suggest that the new techniques used in ChatGPT are helpful in improving the model’s mathematical reasoning abilities.

3.2 Arithmetic and Counting

In this section, we present the evaluation results of ChatGPT’s performance on arithmetic and counting. We first discuss the performance of ChatGPT on arithmetic operations and then move on to its performance on counting tasks.

3.2.1 Arithmetic

We observe that ChatGPT’s accuracy in arithmetic operations declines as the numbers used in the operations increase in size. In particular, the accuracy of multiplication decreases significantly and at a faster rate than addition and subtraction. The accuracy scores for addition and subtraction remained relatively similar. This trend is expected as multiplication is more complex than addition or subtraction, which could explain the larger decrease in accuracy.

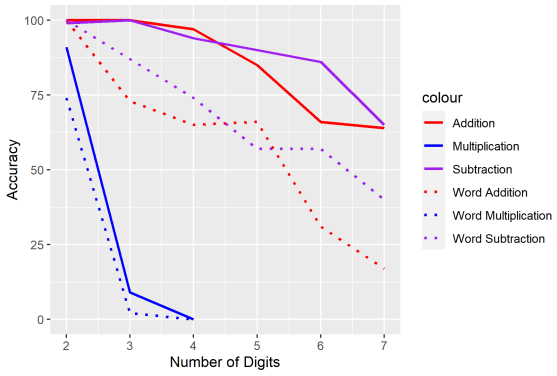

Figure 2: Accuracy of ChatGPT on arithmetic with varying number of digits

3.2.2 Counting

Table 3 shows the accuracy of ChatGPT on counting tasks for different ranges of the number of letters in the input. We observe that the performance of ChatGPT was unexpectedly poor for a relatively simple task. As the length of the input increased, ChatGPT relied on estimation rather than producing an exact answer. For inputs with 50-69 letters, ChatGPT provided the answer ”50” in 66 out of 100 tests.

Number of letters	Accuracy(%)
10-29	22
30-49	9
50-69	3

Table 3: Accuracy on different ranges of counting tasks

In summary, our results indicate that ChatGPT's accuracy in arithmetic operations declines as the numbers used in the operations increase in size, and its accuracy in multiplication is significantly lower than in addition and subtraction. Additionally, ChatGPT's performance on counting tasks was unexpectedly poor, and it relied on estimation rather than producing an exact answer for longer inputs.

3.3 Word Variation

To further test the ability of ChatGPT to synthesize and apply arithmetic rules, we asked the arithmetic questions in the form of English words rather than Arabic numerals. We are motivated by the fact that word variations of these equations are much less likely to appear on the internet, yet contain identical meanings. This category of testing enforces that ChatGPT will not be able to copy information from training, but rather synthesize and apply the rules of arithmetic.

Our results, shown in Figure 2, indicate that the accuracy of ChatGPT in every arithmetic category drops significantly when we use the word variation. This indicates that ChatGPT is reliant on recognizing specific patterns in the input data and reproducing those patterns when answering questions. ChatGPT is not good at synthesizing the rules of arithmetic and applying them in a more general sense. These findings are consistent with previous studies that have shown that large language models such as GPT-3 are not truly "general" in their ability to reason and perform tasks, but rather rely on memorization and pattern recognition [Brown et al., 2020].

4 Error Analysis

In this section, we examine the errors made by ChatGPT and explore potential reasons for these errors. We present specific examples to illustrate the trends we have observed.

4.1 Average Percent Error Analysis

In this section, we provide an overview and analysis of the Average Percent Error (APE) metric used to evaluate the performance of ChatGPT on arithmetic and counting tasks. We explain why a single metric of accuracy may not accurately capture the results of ChatGPT and show APE scores for different tasks in Tables 5, 6, and 4.

Accuracy is a useful metric for determining how precise the answers of a model are, but it only provides a binary classification of correct or incorrect answers. APE, on the other hand, measures how close ChatGPT's answers are to the correct answers, even if they are wrong.

For the arithmetic task, we observe that although the accuracy of multiplication for four digits or higher is 0, the APE scores are around 20%. This indicates that ChatGPT is not completely incapable of performing operations on these large numbers but is rather imprecise. Additionally, a significant portion of the percent error is due to an extra digit. We will discuss this error type in detail in the next section.

The APE scores for the word problems and counting tasks are all less than 20, and some are even below 10. For instance, although ChatGPT's accuracy is below 10 in the counting task for the 30-69 letter range, it's APE score is not very bad. This suggests that ChatGPT has the potential to estimate well, even in challenging tasks where its accuracy is low.

Dataset	APE (%)
AddSub	1.1
SingleEq	7.8
SVAMP	18.7
MultiArith	10.2

Table 4: APE on word problems

Operations	Number of Digits					
	2	3	4	5	6	7
Addition	0%	0%	0%	36.7%	24.9%	13.1%
Subtraction	0%	0%	0.1%	12.6%	3.7%	22.8%
Multiplication	0%	18.3%	20.1%	0.1%	3.4%	10.5%

Table 5: APE of ChatGPT on arithmetic

4.2 Adding One Extra Digit

One common error pattern in the incorrect test cases for large addition, subtraction, and multiplication problems is ChatGPT's tendency to add one

Number of letters	APE (%)
10-29	6.9
30-49	9.4
50-69	18.7

Table 6: APE on counting

Operation	One extra digit error (%)
Addition	18.8
Subtraction	18.8
Multiplication	0

Table 7: Proportion of errors due to an extra digit

extra digit. This error is especially prevalent when the problem requires "carrying the one" or working with large numbers. Table 7 shows that these errors make up 18.8% of the total errors for addition and subtraction. However, this is not prevalent in multiplication as the errors are more than a single extra digit.

To illustrate this error, we present two examples of addition errors where ChatGPT mistakenly added one extra digit in the middle of the number. When prompted with "Respond with only the answer to the following addition expression: 78093+34269," ChatGPT responded with 1123162 while the correct answer was 112362. Similarly, when asked the answer to the expression "56501-38571," it answered with 179330 while the correct answer was 17930.

This deviation from the conventional method of arithmetic calculations suggests that ChatGPT may struggle with longer calculations and maintaining context over the course of the calculation. Further investigation is necessary to understand the underlying causes of this error.

Moreover, these errors may result in inconsistencies when using APE as a metric to evaluate the accuracy of ChatGPT's answers. For instance, an extra digit in the one's place and an extra digit in the thousands place may seem similar but can yield drastically different APE results.

In summary, adding extra digits is a recurring error that ChatGPT makes when solving large addition, subtraction, and multiplication problems. This error could be due to the model's struggle to continually keep track of long calculations. Careful consideration is necessary when evaluating ChatGPT's accuracy using metrics such as APE. Future research may explore methods to mitigate this error and improve the model's performance on deeper reasoning tasks.

5 Conclusion

In recent years, natural language processing (NLP) has seen significant advancements, and ChatGPT has emerged as one of the leading models in this field due to its unique architecture and additional reinforcement learning with human feedback (RLHF). While the model has shown promising results in various NLP tasks, including text generation and summarization, our paper aims to address an important gap in ChatGPT's abilities: mathematical reasoning. Our study evaluates ChatGPT's performance on elementary-level math problems and highlights the need for further research to develop models that can reason effectively about mathematical concepts and solve problems that require arithmetic operations. While our findings suggest that ChatGPT's arithmetic and ability to solve word math problems lag behind its coherency and natural language understanding, we acknowledge that the model's performance is still better than that of previous models in this domain. We also recognize the significant impact of pre-training corpus patterns and specific error types on the model's performance, which requires further exploration.

Furthermore, we emphasize the value of using alternative metrics, such as the Average Percent Error (APE), to assess ChatGPT's performance in mathematical reasoning tasks. Our analysis shows that ChatGPT's accuracy may not always be optimal, but it has the ability to estimate the correct answer. This insight contributes to advancing the development of language models for computational tasks and highlights the need for more comprehensive datasets and evaluation metrics to assess model inference and computational abilities more accurately. In conclusion, while ChatGPT has shown potential in NLP, our analysis indicates that there is still much room for improvement in its mathematical reasoning capabilities. Our study provides important insights into ChatGPT's mathematical and logical reasoning abilities, paving the way for future research to improve the model's performance in this domain.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Cunxiang Wang, Boyuan Zheng, Yuchen Niu, and Yue Zhang. Exploring generalization ability of pretrained language models on arithmetic and logical reasoning. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 758–769. Springer, 2021.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*, 2016.

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

Ali Borji. A categorical archive of chatgpt failures, 2023. URL https://arxiv.org/abs/2302.03494.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597, 2015.

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

terry3041. Pychatgpt. https://github.com/terry3041/pyChatGPT, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Fan Zhou, Haoyu Dong, Qian Liu, Zhoujun Cheng, Shi Han, and Dongmei Zhang. Reflection of thought: Inversely eliciting numerical reasoning in language models via solving linear systems. *arXiv preprint arXiv:2210.05075*, 2022.