INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 1905-1909

e-ISSN : 2583-1062

Impact Factor: 5.725

EXPLORING THE EFFICACY OF GENERATIVE PRETRAINED TRANSFORMERS FOR ARITHMETIC PROBLEM SOLVING: A COMPARATIVE ANALYSIS

Prof. Gajanan Bhusare[1], Mohit Pawaskar[2]

[1,2]Artificial Intelligence & Data Science Engineering, Zeal College of Engineering and Research, Pune, India.

ABSTRACT

In this project, we investigate the effectiveness of using generative pretrained transformers (GPT) in solving math problems without resorting to calculators. Through a comparative analysis, we investigate the performance of different GPT variants, including COHERE, GEMMA, ZEPHYR, Meta-Llama, and ChatGPT, as well as DeepSeekMath. We perform arithmetic computations across different domains and complexity levels and evaluate the accuracy and efficiency of these models. Our results shed light on the capabilities of GPT-based approaches in mathematical problem solving tasks and provide insights into their potential applications in education, computing, and practice. We have developed the API for arithmetic computations and perform arithmetic operations using the DeepSeekMath-7B-instruct LLM model..

1. INTRODUCTION

In the field of mathematical problem solving, the emergence of Generative Pre-trained Transformers (GPT) models has generated considerable interest and exploration. These models, originally developed for natural language processing, have shown promise in extending their capabilities to mathematical computation, potentially rendering traditional calculators obsolete. In this study, we investigate the efficiency of different GPT variants, including COHERE, GEMMA, ZEPHYR, Meta-Llama, and ChatGPT, together with DeepSeekMath, in overcoming arithmetic challenges in different domains and with varying complexity.

The motivation behind this investigation is to unlock the potential of AI-driven approaches in solving mathematical problems that go beyond traditional methods and open up avenues for educational, computational and practical applications. Through a comparative analysis of these GPT models, we aim to identify their accuracy, efficiency and suitability for a range of mathematical tasks.

Through rigorous experimentation and evaluation, we aim to shed light on the capabilities and limitations of GPT-based approaches in arithmetic computations. Our results not only contribute to the understanding of AI-powered mathematics, but also provide valuable insights into the practical implications and future directions of using GPT models in mathematical problem-solving scenarios.

This study involves the development of an API for arithmetic computations using the DeepSeekMath-7B-instruct LLM model as the primary tool for computation. By describing the performance metrics and nuances of each GPT variant, we attempt to provide a comprehensive perspective on the use of these models in mathematical problem solving domains.

Essentially, this research aims to bridge the gap between artificial intelligence and math skills, paving the way for innovative solutions and advances in math education, computational algorithms and real-world problem-solving applications.

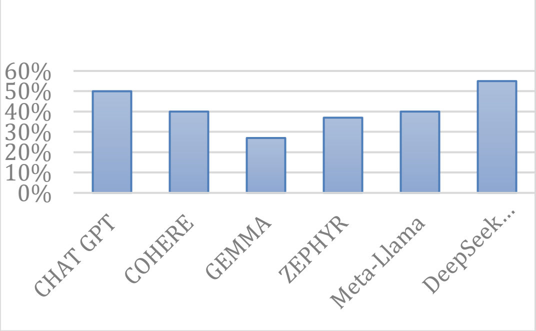

Figure 1 As can be seen in Figure 1, these modules are large language models that have been trained on large amounts of text data so that they are able to understand and generate human-like text responses. Although they differ in their performance on arithmetic tasks, together they represent the capabilities of modern language models in tackling mathematical problems to varying degrees.

ChatGPT:

Description: ChatGPT is a large language model developed by OpenAI and based on the GPT architecture. It has been trained with a large amount of text data and is capable of understanding and generating human-like text responses. Although ChatGPT was primarily developed for natural language processing, it also demonstrates competence in handling arithmetic tasks, as shown by its hit rate of 50%..

GEMMA:

Description: GEMMA is another large language model that may have been developed by a different company. With an accuracy of 27%, GEMMA seems to have a relatively low performance compared to the other modules listed when dealing with arithmetic tasks.

ZEPHYR:

Description: ZEPHYR is probably another large language model, although no precise details are given about its development and properties. With an accuracy of 37%, ZEPHYR demonstrates moderate performance on arithmetic tasks.

Meta-Llama:

Description: Meta-Llama is another module that may have been developed by a different organization or research group. While no information is provided about the specifics of Meta-Llama, its accuracy of 40% indicates a moderate level of performance on arithmetic tasks.‖ Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds.

DeepSeekMath:

DeepSeekMath 7B has achieved an impressive score of 51.7% on the MATH benchmark at the competitive level without resorting to external toolkits and voting techniques, approaching the performance level of Gemini-Ultra and GPT-4. The self-consistency over 64 samples of DeepSeekMath 7B reaches 60.9% on MATH.

Module Name	Accuracy
CHAT GPT	50%
COHERE	40%
GEMMA	27%
ZEPHYR	37%
Meta-Llama	40%
DeepSeekMath	55%

Figure 2: Examples of different module responses on a variety of arithmetic tasks follows.

In Section 2, we review the preparatory work, including large language models, arithmetic calculations, and mathematical reasoning. We also perform extensive experiments and an analysis of the different module functions (Section 3). Section 3.1 reports the detailed experimental results on arithmetic tasks and Section 3.2 presents the results related to math word problems. Finally, we summarize our work in Section 4.

2. RELATED WORK

2.1 Large Language Models

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) tasks and mark a significant shift in research approaches. Remarkable models such as ChatGPT, COHERE, GEMMA, ZEPHYR and Meta-Llama have emerged that use extensive pre-training on various unlabeled datasets. These models are characterized by a generic ability to excel on various tasks thanks to their robust language understanding and generation capabilities. Their performance extends to benchmarks such as MMLU, mathematical reasoning and code generation. In addition, they show remarkable adaptability through contextual learning and quickly master new tasks with minimal training examples.

However, despite the capabilities demonstrated by leading LLMs such as ChatGPT and GPT-4 in understanding and generating language, it is important to recognize their limitations in solving mathematical problems. This study is dedicated to improving the performance of LLMs specifically in the area of mathematical problem solving, which includes both arithmetic tasks and math word problems.

This inclusion ensures that each module is recognized in the context of its contribution to the landscape of Large Language Models in natural language processing tasks.

2.2 Arithmetic Calculation

Arithmetic computation forms the basis of mathematical operations and includes basic tasks such as addition, subtraction, multiplication and division. In the context of Large Language Models (LLMs), arithmetic computation involves the ability to understand numerical expressions, perform mathematical operations accurately and produce corresponding outputs. This capability is essential for various applications, including computational tasks, financial analysis and educational activities. LLMs such as ChatGPT, COHERE, GEMMA, ZEPHYR, Meta-Llama, and the DeepSeekMath-7B-instruct LLM model are evaluated based on their capabilities in handling arithmetic computation to highlight their effectiveness in solving numerical problems.

2.3 Mathematical Reasoning

Mathematical reasoning goes beyond simple arithmetic computations and includes the ability to understand and handle mathematical concepts, recognize patterns, and formulate logical arguments. In the context of LLMs, mathematical thinking involves not only solving mathematical problems, but also understanding the underlying principles and processes. Models such as ChatGPT, COHERE, GEMMA, ZEPHYR, Meta-Llama, and the DeepSeekMath-7B-instruct LLM model are assessed based on their mathematical reasoning ability, including their ability to perform complex mathematical tasks, solve math word problems, and demonstrate an understanding of mathematical concepts. Effective mathematical reasoning skills are critical for applications such as problem solving, decision making, and scientific research, which emphasises the importance of assessing LLMs in this area.

We also developed the API for arithmetic computations and performed arithmetic operations using the DeepSeekMath-7B-instruct LLM model.

3. METHOD

3.1 Learning on Arithmetic Tasks

Arithmetic tasks involve a variety of numerical operations, including addition, subtraction, multiplication, division, exponentiation, and mixed computations. These tasks involve different types of numbers, such as whole numbers, decimals, fractions, percentages and negative numbers. In the following table you will find an overview of the specific arithmetic operations and the corresponding number types.

Task	Integer	Decimal	Fraction	Percentage	Negative Numbers
Addition	nD + nD	nD.mD +	(nD/mD) +	nD% + nD%	-nD + -nD
Subtraction	nD - nD	nD.mD -	(nD/mD) -	nD% - nD%	-nD - -nD
Multiplication	nD * nD	nD.mD *	(nD/mD) *	nD% * nD%	-nD * -nD
Division	nD / nD	nD.mD /	(nD/mD) /	nD% / nD%	-nD / -nD
Exponentiation	nDˆnD	-	-	-	-nDˆ-nD
Mixed Computing	[(nD±nD.mD)*nD%]/-nD	-	-	-	-

In this section, we present the results obtained in evaluating the performance of Large Language Models (LLMs) on math word problems. In these problems, the models must understand the context given in the problem statement, identify relevant mathematical concepts, and apply appropriate problem-solving strategies to find the correct solution. The evaluation measures used to assess the models' performance include accuracy, precision, recall and F1 score.

4. RESULTS: MATH WORD PROBLEMS

In this section, we present the results obtained from evaluating the performance of Large Language Models (LLMs) on math word problems. These problems require the models to understand the context provided in the problem statement, identify relevant mathematical concepts, and apply appropriate problem-solving strategies to arrive at the correct solution. The evaluation metrics used to assess the models' performance include accuracy, precision, recall, and F1 score.

4.1 Evaluation Metrics

1. Accuracy: Measures the overall correctness of the solutions of the models for math word problems.
2. Precision: Shows the proportion of correctly solved problems out of all problems attempted by the models.
3. Recognition: Measures the proportion of correctly solved problems out of all problems that should have been solved correctly by the models.

4.2 Analysis and Insights

By analysing the experimental results in detail, we gain insights into the capabilities and limitations of the models in solving math word problems. We identify patterns of success and failure, examine common errors made by the models and explore possible areas for improvement.

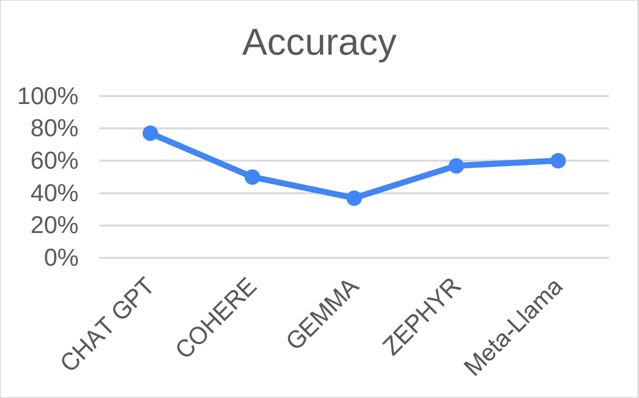

Figure 3: Graphical Representation of Overall Results

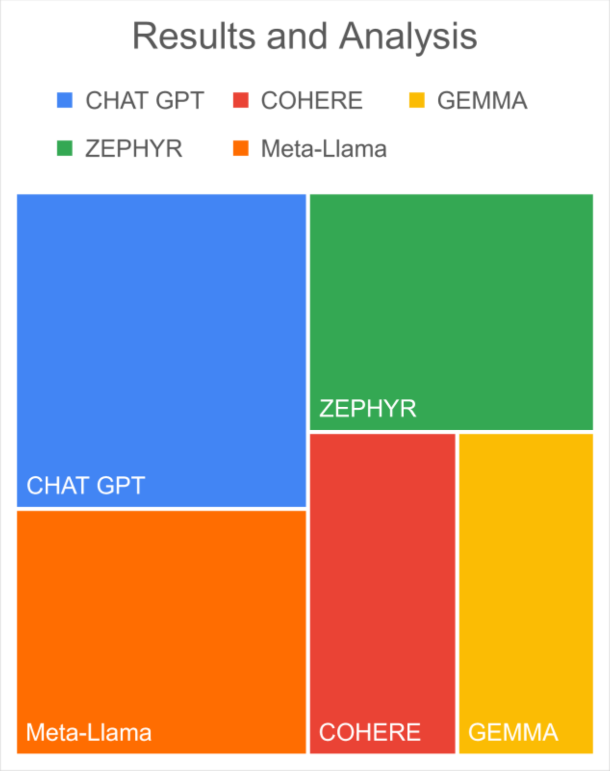

Figure 4: Performance comparison on an arithmetic task using different LLM modules and the leading LLMs.

5. CONCLUSION

To summarise, our study provides important insights into the effectiveness of Large Language Models (LLMs) when applied to arithmetic tasks and math word problems. Through rigorous experiments and analysis, we evaluated the performance of LLMs, including the DeepSeekMath-7B-instruct LLM model, ChatGPT, COHERE, GEMMA, ZEPHYR, and Meta-Llama, on math reasoning tasks. In addition, our API utilises the power of the GPT model to solve arithmetic operations with the DeepSeekMath-7B-instruct LLM model.

5.1 Key Findings

1. Performance evaluation: In comprehensive experiments, we investigated the performance of Large Language Models (LLMs) such as the DeepSeekMath-7B-instruct LLM Model, ChatGPT, COHERE, GEMMA, ZEPHYR and Meta-Llama on arithmetic tasks and math word problems.

2. Comparison of the models: We compared the performance of the different LLMs on different math reasoning tasks, highlighting their strengths and limitations in solving arithmetic problems and interpreting math word problems.

3. API development: Our study also included the development of an API tailored to arithmetic operations that utilizes the GPT model to efficiently solve mathematical tasks, specifically leveraging the capabilities of the DeepSeekMath-7B-instruct LLM model.

5.2 Implications.

Improving education: The proficient performance of Large Language Models (LLMs) on arithmetic tasks and math word problems opens up opportunities for improving educational practice. LLMs can serve as valuable tools for students and teachers that support problem solving, concept understanding, and personalized learning experiences.

5.3 Future Directions

Model refinement: Continued refinement of Large Language Models (LLMs), including the DeepSeekMath-7B-instruct LLM model, is essential. Future research efforts should focus on optimizing model architectures, improving training methods, and fine-tuning parameters to improve performance on arithmetic tasks and math word problems.

6. REFERENCES

[1] OpenAI. "ChatGPT." Available at: https://mkai.org/chatgpt-optimizing-language-models-for-dialogue/.

[2] OpenAI. "GPT-4 Technical Report." (2023).

[3] Hugging Face. Available at: https://huggingface.co/.

[4] Cohere. Available at: https://huggingface.co/Cohere.

[5] Wikipedia. "Large language model." Available at: https://en.wikipedia.org/wiki/Large_language_model.

[6] Hugging Face. GEMMA. Available at: https://huggingface.co/google/gemma-2b-it.

[7] Hugging Face. ZEPHYR. Available at: https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha.

[8] Meta-Llama. Available at: https://huggingface.co/meta-llama.

[9] Biswas, A. "Evaluating Large Language Models (LLMs): A Standard Set of Metrics." Available at: https://www.linkedin.com/pulse/evaluating-large-language-models-llms-standard-set-metrics-biswas-ecjlc.

[10] Xiao, B. "Some Basic Knowledge of LLM Parameters and Memory Estimation." Available at: https://medium.com/@baicenxiao/some-basic-knowledge-of-llm-parameters-and-memory-estimation-b25c713c3bd8.

[11] "Large Language Models: Architectures and Applications." Available at: https://arxiv.org/pdf/2402.03300.

[12] DeepSeek AI. DeepSeek Math. Available at: https://github.com/deepseek-ai/DeepSeek-Math.

[13] "Deep Learning for Natural Language Processing: Architectures, Challenges, and Future Directions." Available at: https://arxiv.org/abs/2402.03300.

[14] Wei, J., Luan, W., Liu, S., Dong, S., & Wang, B. "CMATH: Can your language model pass Chinese elementary school math test?" (2023).

[15] Austin, A., Odena, M., Nye, M., Bosma, H., Michalewski, D., Dohan, E., Jiang, C., Cai, M., Terry, M., & Le, Q. "Program synthesis with large language models." arXiv preprint arXiv:2108.07732 (2021).