Testing LLM's on elementary math problems

Aditya Gupta
SR No: 22205

September 30, 2023

Introduction

The purpose of this assignment is to test the performance of popular large language models on some basic mathematical word problems. The models we will test are ChatGPT, Bard and Bing Chat. The questions used are grade school level and require basic reasoning and some linear algebra to solve. We record the performance of the models based on their accuracy with zero shot, one shot, few shot and chain of thought prompting methods.

ChatGPT

ChatGPT is currently the most popular LLM and is widely used in everyday practices by millions of users. It showed a high degree of accuracy while tackling the problems I gave it from the problem set. The method of solving was detailed and precise, with proper steps and calculations.

However, it failed when presented with questions which also relied on its understanding of the english language in a mathematical context. It would look at the numbers correctly but fail to see some conditions that have been imposed in english. Most commonly, it would miss words like "reduced" or "removed" and continue adding these quantities to the final answer.

An example to demonstrate this failure would be question 9 from the problem set. The question was as follows:

John drives for 3 hours at a speed of 60 mph and then turns around because he realizes he forgot something very important at home. He tries to get home in 4 hours but spends the first 2 hours in standstill traffic. He spends the next half-hour driving at a speed of 30mph, before being able to drive the remaining time of the 4 hours going at 80 mph. How far is he from home at the end of those 4 hours?

ChatGPT could solve for the distances in the three individual steps. However, it failed to understand that the distance from home was the difference between the distance travelled in the first 3 hours and the distance travelled in the last 4 hours. It would add the distances instead of subtracting them. Even after giving multiple prompts, it could not understand the question. Ultimately, I had to instruct it to subtract the distance of the last 4 hours from the distance in the first 3 hours via a prompt.

Bard

Bard is a LLM released by Google. It had the highest zero shot accuracy among the three models being tested. The answers were properly explained and presented in a clear format. The errors could easily be rectified with one or two extra prompts.

It has a better understanding of the importance of english in these mathemamtical word problems. It could capture important details from the prompt and interpret them in a mathematical context. Still, it fell victim to misinterpretation in some basic questions.

Bing Chat

Bing Chat is the LLM made by Microsoft. It uses GPT-4 in its backend, which is considered a very powerful language model. It had a better accuracy than ChatGPT, which works on GPT-3.5.

The model breaks down the prompt into smaller phrases and interprets them individually by trying to perform a relevant web search. This proves harmful for problems where the question is long and the several phrases are closely linked. Although it could figure out the connection, it often made errors in the calculations due to a lack of context.

An example of this would be question 17 from the problem set. The question was as follows:

Two trains leave San Rafael at the same time. They begin traveling westward, both traveling for 80 miles. The next day, they travel northwards, covering 150 miles. What's the distance covered by each train in the two days?

Bing Chat got confused by the wording of the statement and could not understand that the two trains are moving together. It also could not realise that the trains were on the same path, and hence the distance covered is the same for them. The question asks for the distance covered by each train, but it still adds the two distances and gives that as the output. This shows how the model is unable to understand the context of the question and fails to answer it correctly.

Conclusion

We can see that each LLM has its own merits and demerits. Bard is a clear winner here is terms of accuracy. Bing Chat falls short as it fails on longer problems. Still, all models tested had a high accuracy in solving these basic questions. With proper examples and prompts, every model could arrive at the correct answer.

Q No.	Zero shot	One shot	Few shot	Chain of thought
1	✓	✓	✓	✓
2	✓	✓	✓	✓
3		✓		✓
4	✓		✓	
5				✓
6	✓	✓	✓	✓
7	✓	✓	✓	✓
8		✓	✓	
9				
10	✓	✓	✓	✓
11			✓	✓
12	✓	✓	✓	✓
13	✓	✓	✓	✓
14		✓		✓
15	✓	✓	✓	✓
16	✓	✓	✓	✓
17			✓	✓
18	✓	✓	✓	✓
19	✓	✓	✓	✓
20		✓	✓	✓

Table 1: Results obtained by ChatGPT

Q No.	Zero shot	One shot	Few shot	Chain of thought
1	✓	✓	✓	✓
2	✓	✓	✓	✓
3		✓	✓	✓
4	✓	✓	✓	✓
5	✓	✓	✓	✓
6	✓	✓	✓	✓
7	✓	✓	✓	✓
8	✓		✓	✓
9			✓	✓
10	✓	✓	✓	✓
11	✓	✓	✓	✓
12	✓	✓	✓	✓
13	✓	✓	✓	✓
14		✓	✓	✓
15	✓	✓	✓	✓
16	✓	✓	✓	✓
17	✓	✓	✓	✓
18	✓	✓	✓	✓
19	✓	✓	✓	✓
20			✓	✓

Table 2: Results obtained by Bard

Q No.	Zero shot	One shot	Few shot	Chain of thought
1	✓	✓	✓	✓
2	✓	✓	✓	✓
3	✓	✓	✓	✓
4	✓	✓	✓	✓
5		✓	✓	✓
6	✓	✓	✓	✓
7	✓	✓	✓	✓
8			✓	✓
9			✓	✓
10	✓	✓	✓	✓
11	✓	✓	✓	✓
12	✓	✓	✓	✓
13			✓	✓
14	✓	✓	✓	✓
15	✓	✓	✓	✓
16	✓	✓	✓	✓
17				✓
18	✓	✓	✓	✓
19	✓	✓	✓	✓
20	✓	✓	✓	✓

Table 3: Results obtained by Bing Chat