SOLVING AMC QUESTIONS WITH LLMS

Cindy Huang
Columbian High School
Mehta+ Tutoring

Bogdan Kremeznoy
The Brooklyn Latin High School
Mehta+ Tutoring

Hale Türeli
Mehta+ Tutoring

Anhaar Wasi
Mehta+ Tutoring

July 25, 2024

ABSTRACT

AI has improved a lot over the past few years in domains like text summarizations and data categorizations. The aim of this project is to investigate how well the modern AI performs in logical reasoning. Specifically, we assessed the performance of two Large Language Models (LLMs) – Mistral 7B and Llama 3.1 8B – on the American Math Competition (AMC) after fine-tuning these models. For fine-tuning, we used a dataset that contained the AMC problems. We evaluated the models based on the accuracy before fine-tuning and after fine-tuning to see whether we were able to improve upon the existing sources and which model performs better on average. However, as training large language models requires large amounts of computational resources, we did not have access to large amounts of GPU, and so the results are limited. However, we still observe some improvement upon baseline performance when fine-tuning on problem-solving data: for Mistral 7B, the accuracy increased by 4.2% after fine-tuning, and for Llama 3.1 8B, the accuracy increased by 3.7% after fine-tuning. Based on these promising results, we expect that with more computational power, we can achieve even better performance.

1 Introduction

With the rise of LLM models like ChatGPT and Gemini, computers have begun to complete menial tasks previously performed by humans, making the process of completing these tasks cheaper and faster. Despite this, LLMs struggle to solve questions involving competitive mathematics and problem-solving, due to the inability to perform high-level logical reasoning similar to that of human beings. Thus, these models still have a large gap between humans in brain processes. However, recently, LLMs such as Mistral 7B and Llama 3.1 8B seemingly started performing better on math-related tasks, including competition math. Thus, we aim to explore how we can improve these models even more by fine-tuning on mathematics competition problems.

2 Related Work

AMC consists of 25 questions in order of increasing difficulty, with the last 5 problems being extremely difficult. In official testing by the Mistral AI developers, Mistral 7B, LLaMA 2 13B, LLaMA 2 7b, and LLaMA 1 34B were trained on the MATH dataset, which consists of both easy and extremely difficult questions. However, we specifically chose to train our models on problems 1 to 15 of the AMC to make it easier for the models and see whether that affects the accuracy of the fine-tuned models.

3 Methodology

3.1 Dataset

For this project, we utilized the AMIO parsed "Art Of Problem Solving" dataset. This dataset includes problems from AMC (American Mathematics Competition), AIME (American Invitational Mathematics Examination), USAMO (US American Math Olympiad), and some other competitions scrapped from the AoPS (Art Of Problem Solving) website.

3.2 Preprocessing

Since we were specifically working with the AMC problems, we got rid of all the other competitions. Next, since AMC consists of 25 questions in increasing difficulty, with the last problems being extremely difficult to solve. So, we filtered out problems 1 to 15 to make it easier for the models and thus potentially get better results. Afterwards, we split the data into 80% training data and 20% testing data.

The training dataset also includes duplicated problems. However, we decided to leave it this way because for each duplicated problem, there was a different solution. This way, we presented our models with different approaches to solve the same problem.

3.3 Inputting Data Into the Model

Unsloth, which is what we used to fine-tune our models, required the data to be formatted in a certain way. The following shows this format:

	instruction	input	output
71	Solve the math problem below.	Lola, Lolo, Tiya, and Tiyo participated in a p...	In total, there will be $\binom{4}{2} \cdot 2 ...
1968	Solve the math problem below.	You are playing a game. A 2×1 rectang...	We realize that every 2×1 rectangle m...
5484	Solve the math problem below.	A ship sails 10 miles in a straight line fro...	Let C_1 be the point the ship would reach if...
70	Solve the math problem below.	Lola, Lolo, Tiya, and Tiyo participated in a p...	We can calculate the total number of wins ($1...

Figure 1: Training Data Sample for Fine-Tuning with Unsloth

From above, "instruction" directs the model what to do, "input" contains the problem statement, and "output" contains the solution. Note that for the testing data, the "output" section was empty.

4 Models

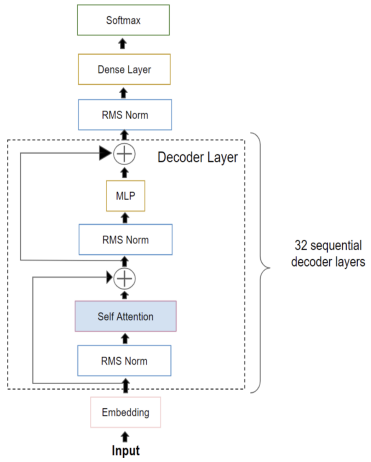
We worked with large language models because they are capable to read and interpret the problem statement and produce a cohesive response. Specifically, we worked with the Mistral 7B Model and LlaMa 3.1 8B Model, two of the best-performing LLMs to date.

4.1 Mistral 7B Model

We chose Mistral 7B Model because it was pretrained for math-related tasks in addition to some other categories. In addition, this model gives high performance performs inference efficiently. The following is the description of Mistral 7B's architecture:[1]

- **Input:** Model receives tokenized text
- **Embedding layer:** Tokens are transformed into vector representations.
- **Decoder Layers:** 32 layers that include:
  - **Self-Attention:** Allows each position to consider the whole input sequence for context.
  - **RMS Norm:** Applied to normalize layer activations, stabilizing the learning process.
  - **MLP (Multi-Layer Perceptron):** Processes the attention output using a gated feed-forward neural network.
- **Dense Layer:** Processed data is shaped into final output format.
- **Softmax:** Gives highest probability of the next token.

4.1.1 Architecture Visualization



Mistral 7B architecture [2]

4.1.2 Output

Estimates a sequence of tokens $\{y_N, ..., y_1\}$:

$$P(y_N, ..., y_1) = \prod_{t=1}^{N} P(y_t | y_{<t})$$

This allows to predict the next token in the sequence of the given tokens.[2]

4.1.3 Activation Function

Mistral 7B uses the Softmax Activation Function:

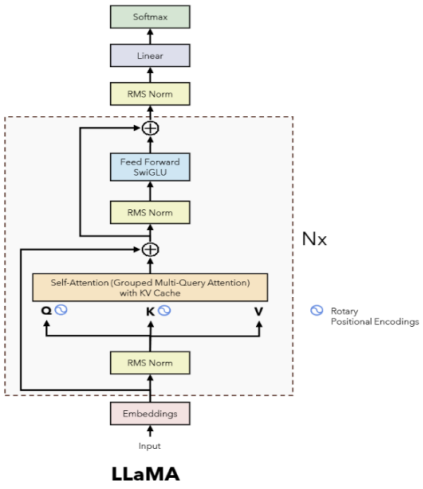
$$f(x) = \frac{e^{xi}}{\sum_j e^{xj}}$$

We utilized this function because it is critical for the output of the model: it gives the highest probability of the next predicted token in the sequence.[2]

4.2 LlaMa 3.1 8B Model

We chose the LlaMa 3.1 8B model because it is a large language model which was not exclusively designed to solve math problems.[3] We wanted to see the difference in performance between a non-math-focused large language model and one that was.

4.2.1 Architecture Visualization



Llama 3.1 8B architecture [4]

5 Results

5.1 Response evaluation of the Mistral 7B Model

Response: **(A)** -20,000
Explanation: $100(100-3)-(100\cdot 100-3) = 100(100-3)-(100\cdot 100-3)$
$= 100(100-3)-(100\cdot 100-3) = 100(100-3)-(100\cdot 100-3) =$

Figure 2: Non-fine-tuned Mistral 7B response

We can use the distributive property to simplify the problem.
$100(100\text{-}3)\text{-}(100\cdot 100-3) = 100\cdot 100-300-10000+3$
$= 10000-300-10000+3$
$= -297$
The answer is -297. So the correct answer choice is C.

Figure 3: Fine-tuned Mistral 7B response

Above, the non-fine-tuned Mistral response got the problem wrong, while the fine-tuned Mistral response got the problem correct.

In Figure 2, the model did not finish its explanation. This is because of the token/compute limitations (and due to the model rambling in its explanation), and so not all responses actually end the explanation. However, we decided to continue using 128 tokens and not increase the size because the model already gave the letter answer in the "Response" section, which is what we wanted. The contents of the model's explanation were also limited: it took the expression "$100(100-3)-(100\cdot 100-3)$" and repeated it four times without making any further logical progress. On the other hand, the fine-tuned Mistral model made meaningful logical steps and got to the correct answer. When we first used 128 tokens for the fine-tuned model, not all the responses given ended with "So the correct answer choice is _" (Figure 3). Instead, some terminated before that, sometimes even mid-sentence as the model was returning "the correct answer choice is ..." Thus, we increased the token size to 256 tokens. which allowed us to get the answer choice for most of the problems. This gave the answer choice for most of the problems, but for some of the responses, the model still didn't finish the sentences.

We also decided to use at most 256 tokens because LLMs suffer from hallucination with longer response windows (i.e. when given more tokens to respond). Therefore, more complicated problems that take longer explanations increase the chance that a response is never returned because the model just starts rambling or repeating itself.

5.2 Accuracy

As mentioned above, not all the responses gave a finished answer, and so the model did not always explicitly return answer choice (e.g., D). So, we had to parse and find if any of the answers were mentioned, so if A was 5, and the model's output mentioned 5, we returned A. However, we filtered out the responses that didn't give the answer either explicitly or implicitly.

Table 1: Accuracy of Mistral 7B and Llama 3.1 8B Before and After Fine-tuning

Model Name	Accuracy
Mistral 7B	13.3%
Fine-tuned Mistral 7B	17.5%
Llama 3.1 8B	11.7%
Fine-tuned Llama 3.1 8B	15.4%

As seen above, the accuracy increased for both models after fine-tuning. However, one may accurately note that these results are insignificant because they are below the percentage if someone would randomly guess (i.e. 20%). However, the AMC is a very tricky competition, and each year, all the problems include new combinations of concepts and ideas. Due to this, the AMC is extremely difficult for anyone without experience in problem-solving, and as a result, it is common to see below 20% of problems being solved correctly.. We also see that in both comparisons of non-fine-tuned and fine-tuned models, Mistral 7B (more math-specific model) outperforms Llama 3.1 8B (not math-specific model).

6 Other challenges

We could not fine-tune the models for a large number of epochs because we were running out of GPU in Google Colaboratory, which is what allows us to train the models in a reasonable amount of time. So, we resorted to using either one epoch of 261 steps or less for training of the Mistral 7B and Llama 3.1 8B models.

7 Conclusion

In our project, we were able to demonstrate that fine-tuning on problems 1 to 15 of the AMC indeed slightly increased the accuracy of both models on these kinds of problems, while Mistral 7B still slightly outperformed Llama 3.1 8B. Thus, we conclude that training the LLMs on easier competition math problems can make the AI actually learn, progress, and be ready to jump to the next difficulty level.

8 Future Work

If we had more time and GPU access, we would train our model for a much larger number of epochs, as it promises to boost the accuracy even more from what we have achieved.

We would also like to improve upon extracting the letter-answer from the fine-tuned model. For our project, the solutions in our training data (in the "output" section of the dataset) ended with "So, the correct answer choice is _." This caused the fine-tuned model (Figure 3) to also give the correct letter-answer at the end and in the same sentence format. However, not all the responses ended in this because of the token limitations, and so we had to play around with the token size to get a valid response for most of the problems. Thus, we want to make this part better by adding something like "The correct answer choice is _" in the beginning of the solution in the training dataset, which would allow us to get the letter-answer for all the responses. Additionally, we want to ensure that the model always gives a completed response by giving the model its previous response and the prompt to generate the next response.

Finally, we believe this project has a potential to extend upon more challenging competitions, like USAMO (US American Mathematics Olympiad), and even IMO (International Mathematics Olympiad), after the model received adequate training on easier competitions and obtained all the basics of math and problem-solving.

9 Division of Labor

We divided the work as follows:

- Data Preprocessing: Bogdan and Anhaar
- Mistral 7B code: Bogdan (most part), Cindy, Hale
- Llama 3.1 8B code: Bogdan
- Paper: Bogdan (most part), Cindy, Hale
- Poster: Bogdan, Cindy, Hale (equally distributed)

10 Acknowledgements

We would like to acknowledge Haripriya Mehta and Bagirath Mehta. Without their constant help, support, feedback, and enthusiasm, this project would not have been possible.

References

[1] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[2] Hamza FILALI BABA. Mistral: Decoding the complexities of a large autoregressive language model. *hamza-onai.com*, Dec 2023.

[3] AI @ Meta Llama Team. The llama 3 herd of models. *Meta*, Jul 2024.

[4] vignesh yaadav. Exploring and building the llama 3 architecture : A deep dive into components, coding, and inference techniques. *Medium*, Apr 2024.