

# Team QED



Adnan Cihan Cakar, Matthew Gelvin, Cisil Karaguzel, Hatice Mutlu, Ming Zhang

The Erdős Institute  
Data Science Bootcamp

## Introduction:

- The state-of-the-art language models have achieved human-level performance on many tasks but still face significant challenges in multi-step mathematical reasoning.
- Recent advancements in large language models (LLMs) have demonstrated exceptional capabilities across diverse tasks, including common-sense reasoning, question answering, and summarization.
- However, they struggle with tasks requiring quantitative reasoning, such as solving complex mathematical problems.
- Mathematics serves as a valuable testbed in machine learning for problem-solving abilities, highlighting the need for more robust models capable of multi-step reasoning.

# Objective:

The primary goal of this project is to develop a customized LLM that can provide step-by-step solutions to math problems by fine-tuning a base LLM using a large mathematical dataset. The major priorities of the project are:

- Fine-tuning the model to handle mathematical notation and reasoning.
- Improving accuracy of the numerical solution compared to the base model.

# Dataset:

## Training Data:

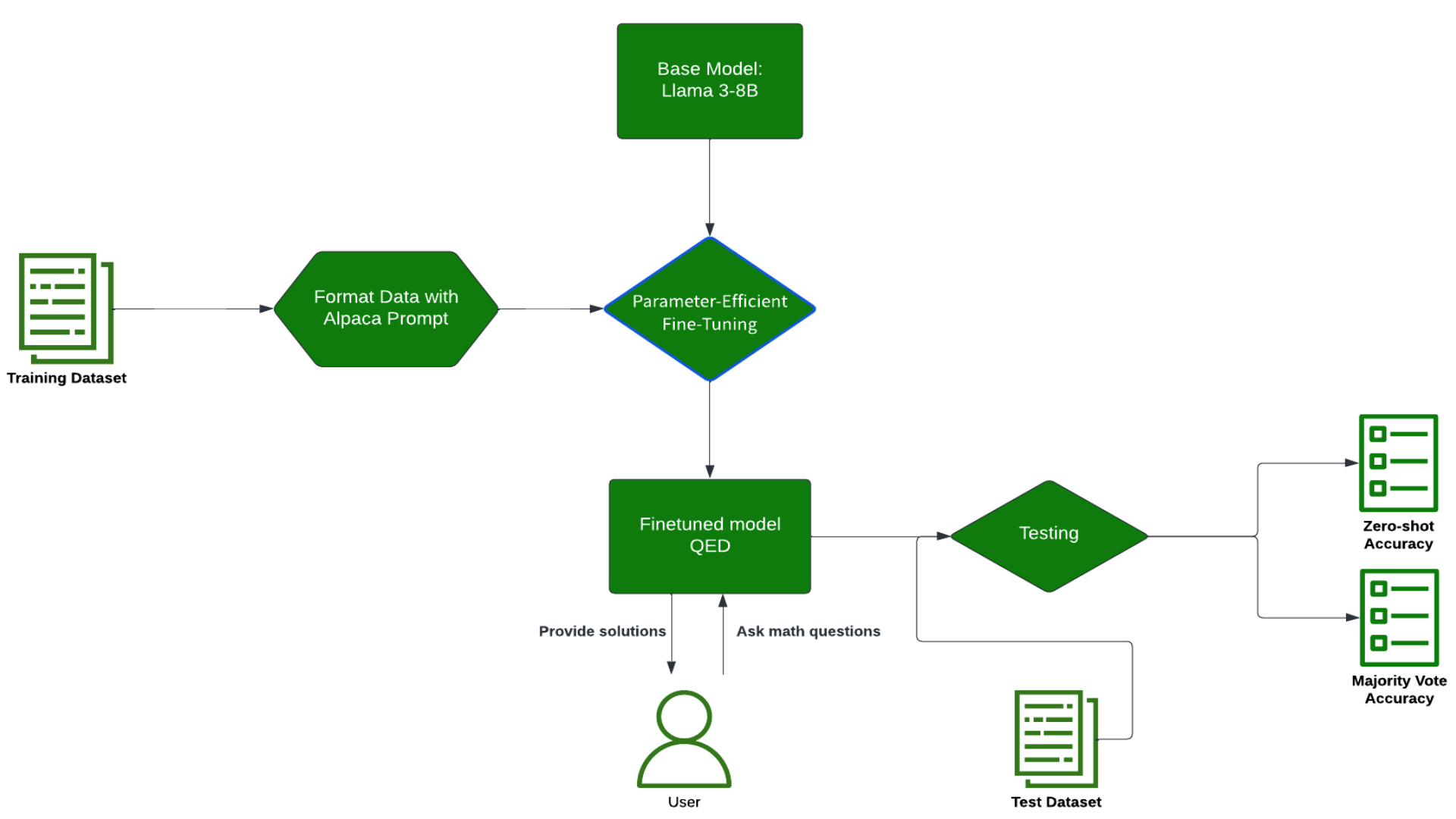
- The [MetaMathQA](#) 🤗: 395k math questions augmented from the **training portions** of the [GSM8K](#) and [MATH](#) datasets:
  - The GSM8k (training part only): grade school level math problems requiring 2 to 8 steps to solve.
  - The MATH (training part only): more challenging problems from prestigious math competitions such as AMC 10, AMC 12, and AIME.

## Test Data:

- The test portions of [GSM8k](#) and [MATH](#)

## Sample data:

```
{ "question": "What is the positive difference between $120\%$ of 30 and $130\%$ of 20?",  
  "solution": "One hundred twenty percent of 30 is $120\cdot30\cdot\frac{1}{100}=36$, and  
$130\%$ of 20 is $130\cdot20\cdot\frac{1}{100}=26$. The difference between 36 and 26  
is $\boxed{10}$.",  
  "answer": "10" }
```



# Model and Training Procedure:

- Using the Alpaca prompt to format our dataset.
- Starting with Unsloth's [llama-3-8b-bnb-4bit](#) model, a quantized 4-bit model using bitsandbytes, which allows for faster fine-tuning and reduced memory usage. It is a small model with 4.65B parameters.
- Adding LoRA adapters to update only 1-10% of parameters, allowing faster fine-tuning.
- Fine-tuning the base model with supervised instruction-response pairs using Hugging Face's SFFT trainer for one epoch.
- The model was trained with A100 GPU for 16 hours.

# Model Testing:

- We performed a hyperparameter search on the temperature of the model, which controls the variance of the output.
- Performance was judged by autoregressively sampling a single low-temperature solution, extracting and checking if the final answer is correct.
- The evaluation datasets include **the test portions** of GSM8k and MATH, which are independent of the training dataset and ideal for assessing our model's generalizability.
- When computing the accuracy of our model, the evaluation functions (extracting, parsing, and comparing math quantities) developed by the [Xwin-Math team](#) were used.
- Solution generation with and without majority voting were both tested.

# Results:

Model	GSM8k dataset	MATH dataset
<code>unsloth/llama-3-8b-bnb-4bit</code> (Base Model)	Accuracy: 29.4 %	Accuracy: 7.0 %
<code>acihanckr/erdos_qed_2024</code> (Our Model)	Accuracy: 35.2 %	Accuracy: 7.8 %



# Majority Vote:

- We evaluated our model's performance using majority voting on 50 randomly selected questions from the GSM8k test dataset, generating 5 high temperature solutions per question.

Our model	Accuracy: 44 %
Our model with majority vote	Accuracy: 46 %

# Sample Solution:

- *Question:*

Jane planted a beanstalk in her backyard. After the first week, it was 3 inches tall. It doubled in height the second week. It grew another 4 inches in the third week. How tall was the beanstalk after 3 weeks?

- *Generated solution:*

Let's think step by steps. The first week, the beanstalk was 3 inches tall. The second week, it doubled in height, so it was 6 inches tall. The third week, it grew another 4 inches, so it was 10 inches tall. So, after 3 weeks, the beanstalk was 10 inches tall.

# Conclusion:

- We efficiently fine-tuned the Llama 3-8B model on a large dataset of math problems, achieving a substantial performance boost relative to the baseline.
- Despite this improvement, our model's performance still falls short of the latest state-of-the-art models. Specifically, while we observed notable gains in accuracy, especially on the GSM8k dataset, there remains significant room for improvement in solving more complex mathematical problems as evidenced by our results on the MATH dataset.
- Further refinement and advanced techniques are necessary to bridge the gap and achieve superior performance in mathematical reasoning tasks.

Comparison to the other models

Model	GSM8k Pass@1	MATH Pass@1
MPT-7B	6.8	3.0
Falcon-7B	6.8	2.3
LLaMA-1-7B	11.0	2.9
LLaMA-2-7B	14.6	2.5
MPT-30B	15.2	3.1
LLaMA-1-13B	17.8	3.9
GPT-Neo-2.7B	19.5	--
Falcon-40B	19.6	2.5
Baichuan-chat-13B	23.9	--
Vicuna-v1.3-13B	27.6	--
LLaMA-2-13B	28.7	3.9
InternLM-7B	31.2	--
ChatGLM-2-6B	32.4	--
GPT-J-6B	34.9	--
LLaMA-1-33B	35.6	3.9
LLaMA-2-34B	42.2	6.24
RFT-7B	50.3	--
LLaMA-1-65B	50.9	10.6
Qwen-7B	51.6	--
WizardMath-7B	54.9	10.7
LLaMA-2-70B	56.8	13.5

(Source: MetaMath)

# Further Directions:

- Training a larger base model (e.g. Llama 3-70B)
- Training for two to three epochs on a larger math problem dataset.
- Training an LLM as a verifier to judge the correctness of model-generated solutions, aiming to reduce false positives and improve overall accuracy.

# References:

1. <https://huggingface.co/unsloth/llama-3-8b-bnb-4bit>
2. <https://github.com/Xwin-LM/Xwin-LM/tree/main/Xwin-Math>
3. <https://huggingface.co/datasets/meta-math/MetaMathQA>
4. <https://github.com/openai/grade-school-math>
5. <https://github.com/hendrycks/math>

# Thank you!

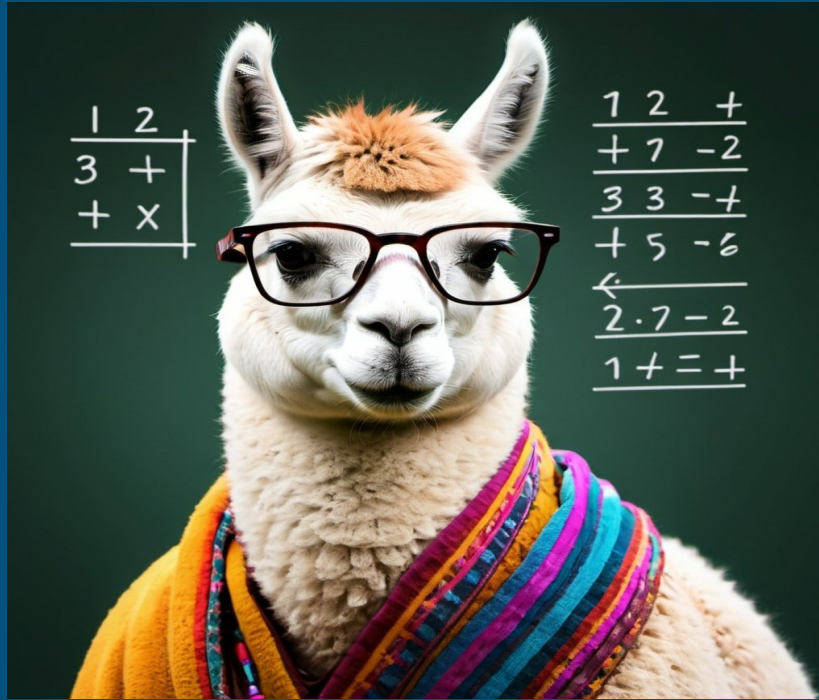


Photo is generated by openart.ai.