**Group Members:** Yasir Raza, Umair Tufail, Muhammad Talha Abdullah, Ali Amjad

**Group Leader:**  Yasir Raza

**Project Overview:**

Converting LATEX expressions into Python code through LLMS

**Steps:**

**1 – Synthetic Data Generation:**

Synthetic dataset was generated using sympy and llama 70B that is based on 8000 mathematical expressions of 14 types including algebraic, logarithmic and polynomial types etc.

Difficult ones like Differentiation, Integration, Derivatives were given a larger part in the dataset.

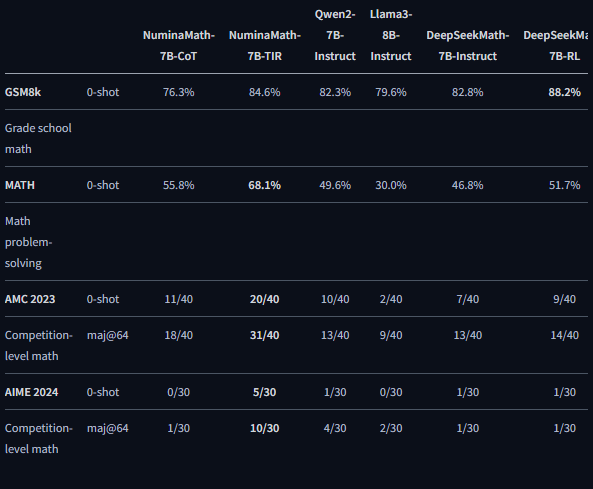
**2 – Converting:**

We converted this data in both “text book” and “Json” format to evaluate which format is more suitable for our model and “JSON” format yielded the better results.

**3 Choosing the Model:**

We choose our main LLM model “[**NuminaMath-7B-TIR**](https://huggingface.co/AI-MO/NuminaMath-7B-TIR)” which is a fine tuned version of “[**deepseek-math-7b-base**](https://huggingface.co/deepseek-ai/deepseek-math-7b-base)” which was also nominated as no 1 spot in AI math’s Olympiad 2024 outperforming LLMs like GPT and LLAMA due to its main feature **Self-Consistency with Tool-Integrated Reasoning (SC-TIR)**.

This model was already performing pretty well on the go with just inference without training dataset. We understood that this was a machine translation task and Encoder-Decoder Transformers should be our go to. But their inference specially without finetuning was not up to the mark hence we choose this over them. Below are benchmarks showcasing the performance of this model.



At first, we choose phi because of its performance with “Text book” format. Then we tried T5 due to encoding and decoding capabilities but the results were not up to the mark.

aimo\_lora\_v3 was used for training purposes. And Prompt engineered the model to derive the outcomes required. This model was not quantized full-fledged model was utilized.

**Problems Faced:**

* One of the problems we faced was validation of data produced by LLM.
* Post processing also created null values because of the structure of training data’s code and the way it was handling (negative logs)

**References:**

[AI Mathematical Olympiad - Progress Prize 1 | Kaggle](https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize/discussion/519303)