ICLR
International Conference On Learning Representations

OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data


Shubham Toshniwal


Wei Du


Ivan Moshkov


Branislav Kisacanin


Alexan Ayrapetyan


Igor Gitman

NVIDIA

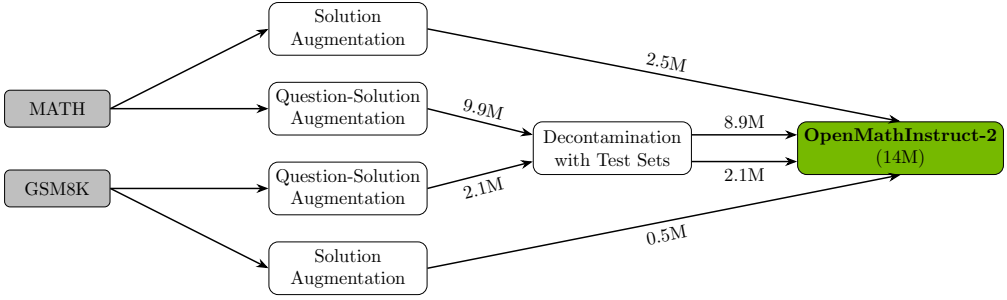
Introduction

State-of-the-art progress in mathematical reasoning has largely remained closed-source due to the lack of access to training data.

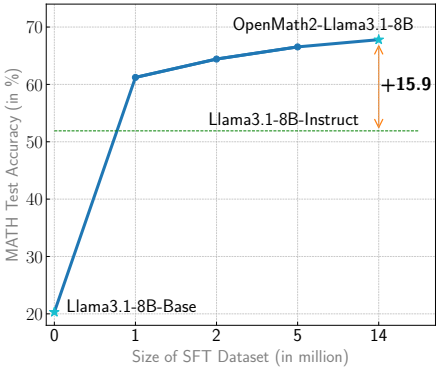
We conducted careful ablations and released OpenMathInstruct-2, a high-quality finetuning (SFT) dataset for math reasoning which consists of 14M question-solution pairs ($\approx$ 600K unique questions).

Finetuning the `Llama-3.1-8B-Base` using OpenMathInstruct-2 outperforms `Llama3.1-8B-Instruct` on MATH by an absolute 15.9% (51.9% $\rightarrow$ 67.8%).

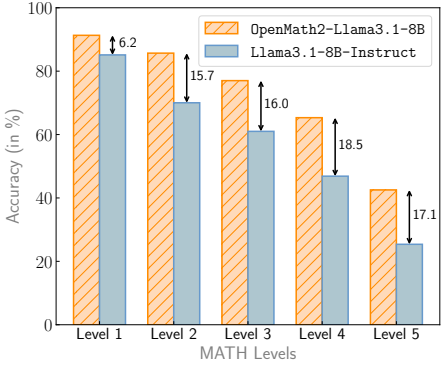
OpenMathInstruct-2: The Complete Pipeline



Impact of OpenMathInstruct-2



Performance of `Llama3.1-8B-Base` on MATH after finetuning on increasing proportions of `OpenMathInstruct-2`.

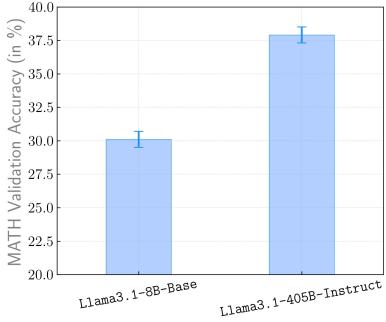


Comparison of `OpenMath2-Llama3.1-8B` and `Llama3.1-8B-Instruct` on accuracy across MATH difficulty levels.

Key Findings I: Choice of Teacher Model

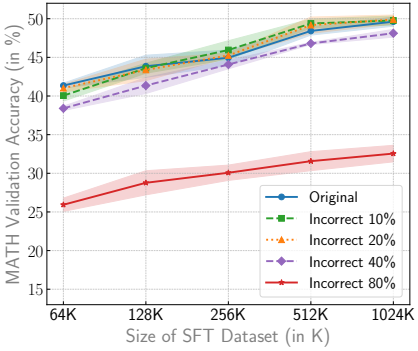
Suppose we have a teacher LLM (M_T) and a student LLM (M_S)

Given an SFT dataset of equal size in both the set of questions and the number of solutions per question, which model's solutions are better suited for teaching a student model?

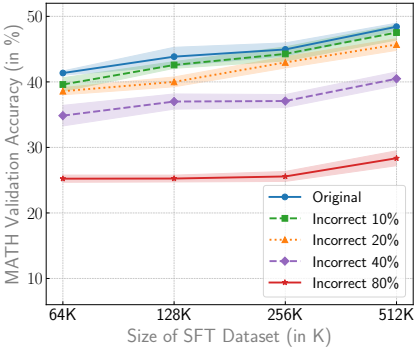


Student model distilled from a stronger teacher model outperforms its self-distilled counterpart

Key Findings II: Robustness of SFT



Correct solutions mismatched with questions



Adding wrong-answer solutions

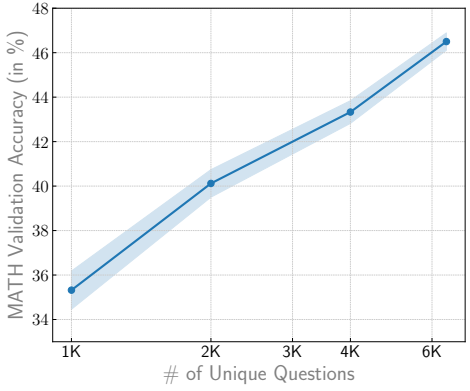
Conducted extensive experiments on removing and adding noisy data to the SFT blend

SFT performance is robust to presence of up-to 20% low-quality data

Key Findings III: Question Diversity

Controlling for the SFT size, how much does the question diversity matter?

Strong positive correlation between # of questions and model performance



Conclusion

We conducted careful ablations to reveal key insights on creating SFT data for mathematical reasoning

Guided via these ablations, we created and released OpenMathInstruct-2 which consists of 14M question-solution pairs ($\approx$ 600K unique questions).

Dataset quality is illustrated by strong finetuning performance. In particular, finetuning the `Llama-3.1-8B-Base` using OpenMathInstruct-2 outperforms `Llama3.1-8B-Instruct` on MATH by an absolute 15.9% (51.9% $\rightarrow$ 67.8%).

The dataset comes with a commercially permissive license, and has been downloaded more than 50K times via Huggingface