StackMathQA: A Curated Collection of 2 Million Mathematical Questions and Answers Sourced from Stack Exchange

Yifan Zhang

ASI Research

yifanzhangresearch@gmail.com

Abstract

The development of sophisticated mathematical reasoning in large language models (LLMs) is often hindered by the scarcity of large-scale, high-quality, and domain-specific training data. To address this gap, we introduce **StackMathQA**, a comprehensive dataset containing nearly 2 million question-and-answer pairs sourced from the Stack Exchange network. This dataset aggregates expert-level and enthusiast discussions from premier platforms, including Math Stack Exchange, MathOverflow, Statistics Stack Exchange, and Physics Stack Exchange. We provide the data in multiple formats and curated subsets created through importance resampling (Xie et al., 2023) to cater to a wide range of research needs, from large-scale pre-training to targeted fine-tuning. This report details the dataset's construction methodology, structure, content, and potential applications, establishing StackMathQA as a valuable resource for advancing machine reasoning in quantitative domains.

`Date:` Released: January 11, 2024

`Dataset:` https://huggingface.co/datasets/math-ai/StackMathQA

1 Introduction

The ability to perform complex mathematical reasoning is a key frontier in artificial intelligence research. While large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding, their proficiency in mathematics often lags, primarily due to the lack of extensive, high-quality training data that mirrors the complexity and diversity of human mathematical discourse (Hendrycks et al., 2021).

To bridge this gap, we present **StackMathQA**, a new large-scale dataset designed to facilitate the training and evaluation of LLMs on mathematical tasks. StackMathQA consists of approximately 2 million question-and-answer (Q&A) pairs meticulously extracted from several high-authority communities within the Stack Exchange network. These platforms are rich with nuanced questions, detailed explanations, and formal mathematical notation, making them an ideal source for training sophisticated reasoning models.

This report provides a comprehensive overview of the StackMathQA dataset. We detail the data sourcing and curation process, the various formats and subsets provided, the structure of the data, and its potential applications in AI research. By releasing StackMathQA, we aim to provide the research community with a robust and versatile resource to accelerate progress in automated mathematical reasoning.

2 Dataset Curation and Preprocessing

The construction of StackMathQA involved a multi-stage process of data sourcing, extraction, and preprocessing to ensure quality and usability.

2.1 Data Sources

The dataset is aggregated from four highly respected Stack Exchange sites, chosen for their focus on quantitative reasoning and the high quality of their community-contributed content:

- **Mathematics Stack Exchange**: A Q&A site for people studying math at any level.
- **MathOverflow**: A Q&A site for professional mathematicians.
- **Statistics Stack Exchange (Cross Validated)**: A Q&A site for people interested in statistics, machine learning, and data analysis.
- **Physics Stack Exchange**: A Q&A site for active researchers, academics, and students of physics.

All content, including questions and answers, was preserved in its original format, retaining the rich LaTeX encoding used for mathematical notation.

2.2 Initial Preprocessing and Formatting

The raw extracted data was processed into two primary formats to offer flexibility for different use cases.

2.2.1 Question-to-Answer-List Format ('stackmathqafull-qalist')

In this format, each entry corresponds to a single question and contains a list of all its associated answers. This structure is useful for tasks that require an understanding of multiple perspectives or solutions for a single problem.

- **Structure**: `"Q": "question", "A_List": ["answer1", "answer2", ...], "meta": ...`

2.2.2 Question-to-Single-Answer Format ('stackmathqafull-1q1a')

Here, each entry is a direct pair of one question and one answer. Questions with multiple answers are duplicated for each answer, creating a flat structure ideal for supervised fine-tuning.

- **Structure**: `"Q": "question", "A": "answer", "meta": ...`

Table 1 summarizes the total number of questions and answers extracted from each source site.

Table 1 Statistics of the fully preprocessed StackMathQA dataset.

Source Website	Unique Questions	Total Answers
math.stackexchange.com	827,439	1,407,739
mathoverflow.net	90,645	166,592
stats.stackexchange.com	103,024	156,143
physics.stackexchange.com	117,318	226,532
Total	**1,138,426**	**1,957,006**

3 Curated Subsets via Importance Resampling

To make the dataset more accessible and cater to research environments with varying computational resources, we created several smaller, high-quality subsets. These subsets were generated using importance resampling from the 'stackmathqafull-1q1a' collection, rather than simple random sampling. This ensures that the smaller datasets retain a high concentration of valuable and informative Q&A pairs, prioritized by community engagement metrics such as answer scores and question view counts.

We provide five curated subsets of decreasing size:

- **StackMathQA1600K**: 1.6 million Q&A pairs.
- **StackMathQA800K**: 800,000 Q&A pairs.
- **StackMathQA400K**: 400,000 Q&A pairs.
- **StackMathQA200K**: 200,000 Q&A pairs.
- **StackMathQA100K**: 100,000 Q&A pairs.

The distribution of Q&A pairs from each source across these curated subsets is detailed in Table 2. As shown, the resampling strategy tends to favor content from the highly active 'math.stackexchange.com community, especially in the smaller subsets.

Table 2 Distribution of Q&A pairs across curated StackMathQA subsets.

Source Website	1600K	800K	400K	200K	100K
math.stackexchange.com	1,244,887	738,850	392,940	197,792	99,013
mathoverflow.net	110,041	24,276	3,963	1,367	626
stats.stackexchange.com	99,878	15,046	1,637	423	182
physics.stackexchange.com	145,194	21,828	1,460	418	179
Total	**1,600,000**	**800,000**	**400,000**	**200,000**	**100,000**

4 Dataset Structure and Usage

4.1 Data Fields

Each entry in the curated datasets (‘stackmathqa100k‘ through ‘stackmathqa1600k‘) and the ‘1q1a‘ format contains the following fields:

- `Q`: (string) The mathematical question, with LaTeX encoded for formulas.
- `A`: (string) The corresponding answer, also with LaTeX encoding.
- `meta`: (dict) A dictionary containing metadata, such as the source URL, question ID, answer score, and other relevant information.

4.2 Loading the Dataset

The StackMathQA dataset is hosted on the Hugging Face Hub and can be easily loaded using the ‘datasets‘ library. Users can select a specific configuration corresponding to the desired subset.

```
from datasets import load_dataset

# Load the default (1.6M) configuration
ds = load_dataset("math-ai/StackMathQA", "stackmathqa1600k")

# To load a different subset, specify its name
# ds_100k = load_dataset("math-ai/StackMathQA", "stackmathqa100k")

print(ds['train'][0])
```

Listing 1 Loading a specific configuration of StackMathQA.

5 Potential Applications

StackMathQA is a versatile resource that can support a wide range of research directions in AI and machine learning:

- **Continual Pre-training**: The large scale of the dataset makes it an excellent resource for continual pre-training of foundation models to enhance their understanding of mathematical language, symbols, and reasoning structures.
- **Supervised Fine-Tuning (SFT)**: The curated Q&A pairs are ideal for fine-tuning LLMs to improve their ability to follow instructions and generate accurate, step-by-step solutions to mathematical problems.
- **Domain-Specific Model Development**: Researchers can use StackMathQA to train or specialize models for expert domains like theoretical physics, advanced mathematics, or econometrics.
- **Benchmark for Mathematical Reasoning**: The dataset can serve as a challenging benchmark to evaluate the performance of LLMs on a diverse set of real-world mathematical queries.

6 Conclusion

We have introduced StackMathQA, a large-scale dataset of nearly 2 million mathematical questions and answers sourced from the Stack Exchange network. By providing meticulously curated and preprocessed data in various formats and sizes, we aim to equip the research community with a powerful tool for advancing the state of the art in machine reasoning. We believe StackMathQA will be instrumental in the development of next-generation language models with robust and reliable mathematical capabilities.

The dataset is publicly available under a Creative Commons Attribution 4.0 International license. We encourage its use and welcome feedback from the community.

References

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.