
MathScale: Scaling Instruction Tuning for Mathematical Reasoning

Zhengyang Tang^{1,2,3} Xingxing Zhang² Benyou Wang^{1,3} Furu Wei²

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in problem-solving. However, their proficiency in solving mathematical problems remains inadequate. We propose *MathScale*, a simple and scalable method to create high-quality mathematical reasoning data using frontier LLMs (e.g., GPT-3.5). Inspired by the cognitive mechanism in human mathematical learning, it first extracts topics and knowledge points from seed math questions and then build a concept graph, which is subsequently used to generate new math questions. *MathScale* exhibits effective scalability along the size axis of the math dataset that we generate. As a result, we create a mathematical reasoning dataset (MathScaleQA) containing two million math question-answer pairs. To evaluate mathematical reasoning abilities of LLMs comprehensively, we construct MWPBENCH, a benchmark of Math Word Problems, which is a collection of 9 datasets (including GSM8K and MATH) covering K-12, college, and competition level math problems. We apply MathScaleQA to fine-tune open-source LLMs (e.g., LLaMA-2 and Mistral), resulting in significantly improved capabilities in mathematical reasoning. Evaluated on MWPBENCH, MathScale-7B achieves state-of-the-art performance across all datasets, surpassing its best peers of equivalent size by 42.8% in micro average accuracy and 43.6% in macro average accuracy, respectively.

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in problem-solving. However, their pro-

iciency in solving mathematical problems remains inadequate, potentially due to the inherent necessity for multi-step complex reasoning in mathematical problem-solving. Instruction Tuning (Wei et al., 2021) is an effective approach to unlock certain capabilities in LLMs. Unfortunately, this approach is constrained by the limited size of the currently available datasets on mathematical reasoning. For example, the most popular math datasets, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b), each only contains around 7.5K training examples.

An effective method to tackle this challenge is to augment existing high-quality math datasets using frontier LLMs such as GPT-3.5 and GPT-4. For instance, WizardMath (Luo et al., 2023) introduces an array of operations for GPT-3.5 to generate math questions with increased complexity. MetaMath (Yu et al., 2023) bootstraps questions in GSM8K and MATH through answer augmentation, question rephrasing, self-verification and FOBAR questions. The newly generated examples by these methods exhibit substantial similarity to the original examples contained within the training set, which limits their power in generating large scale math datasets.

We therefore propose a conceptually simple and scalable method *MathScale*, which is less dependent on original training examples. Specifically, we first prompt GPT-3.5 to extract high level concepts (i.e., topics and knowledge points) from existing seed math questions. In this step, we convert concrete math questions to extractions and the dependency to original questions is largely removed. Given these extractions, we then build a concept graph, which is used to estimate the connections between different concepts. Finally, we can instruct GPT-3.5 to generate new math questions based on randomly sampled concepts from the graph. Intuitively, we can generate significantly more examples using different combination of concepts than using augmentation-based methods, since the resulting number of new examples is bounded by the number of augmentation operations. *MathScale* also bears resemblance to the cognitive mechanisms underlying the process of mathematical learning in humans (Tall, 2013). Tall (2013) argues that the learning process of human involves two distinct steps called *concept compression* and *connection forging*. *Concept compression* mirrors the process of high level concept extraction, while *connection forging* is similar to our

¹The Chinese University of Hong Kong, Shenzhen, China

²Microsoft Research Asia, Beijing, China ³Shenzhen Research Institute of Big Data, Shenzhen, China. Correspondence to: Xingxing Zhang <xingxing.zhang@microsoft.com>, Benyou Wang <wangbenyou@cuhk.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

concept graph construction.

Mathematical capability evaluation is another issue arising from the lack of high-quality mathematical datasets. Recently, most LLMs employ GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) for evaluation. However, GSM8K focuses on elementary-level problems, while MATH offers competition-level challenges. There is a clear gap between the two kinds of capabilities measured. Therefore, we introduce MWPBENCH, a comprehensive and unified benchmark to measure mathematical reasoning capabilities. MWPBENCH is composed of 9 different math word problem datasets (including GSM8K and MATH) and it covers math word problems from elementary school to college level with different difficulty levels. Moreover, MWPBENCH standardizes evaluations across all datasets with a unified protocol, promoting consistent and fair model comparisons.

MathScale exhibits effective scalability along the size axis of the math dataset that we generate. As a result, we create a mathematical reasoning dataset (MathScaleQA) containing two million math question-answer pairs. We apply MathScaleQA to fine-tune open-source LLMs (e.g., LLaMA-2 and Mistral), resulting in significantly improved capabilities in mathematical reasoning. Evaluated on MWPBENCH, MathScale-7B achieves 35.2% in micro average accuracy and 38.2% in macro accuracy, outperforming its best peers of equivalent size by 42.8% and 43.6%, respectively.

2. MWPBENCH Evaluation Framework

2.1. MWPBENCH

Existing Datasets Our first endeavor is to collate established datasets, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), TAL-SCQ (TAL, 2023), Math23k (Wang et al., 2017), Ape210k (Zhao et al., 2020), GaokaoBench-Math (Zhang et al., 2023), and AGIEval (Zhong et al., 2023) series (see Table 1). Types of problems of these datasets are different. For example, most datasets contain math word problems, while TAL-SCQ comprises multi-choice questions. Intuitively, multi-choice questions are simpler because LLMs only need to figure out which choice leads to a higher probability. Therefore, we convert all multi-choice questions to math word problems (detailed in Appendix A.1). Secondly, some of the datasets (e.g., Math23k, Ape210k) are not in English and we translate them to English to expand existing math datasets (detailed in Appendix A.2). Note that we translated part of their training sets and full test sets into English. We also noticed another dataset, Lila (Mishra et al., 2022), which includes math question with code solutions as well as code questions from MBPP (Austin et al., 2021) and APPs (Hendrycks et al., 2021a), focusing differently from ours.

CollegeMath Existing datasets does not cover college-level mathematics which requires diverse skills such as analytical thinking, logical reasoning, and quantitative analysis. We therefore propose CollegeMath to bridge this gap.

We curated a collection of nine college mathematics textbooks, each addressing a distinct topic (see Table 2 for more details). These textbooks encompass seven critical mathematical disciplines: algebra, pre-calculus, calculus, vector calculus, probability, linear algebra, and differential equations. These textbooks are originally in PDF format and we convert them to text format using the Mathpix API¹, where equations are transformed to LaTeX format. Once converted a textbook to text format, we are ready to extract exercises and their solutions. For each book, we first manually segment the book into chapter and identify pages with exercises and their solutions. Then we extract questions in exercises and their associated short answers (see more details of our prompts in Appendix A.3). In total, this dataset contains 1281 examples for training and 2818 examples for test.

2.2. Unified Evaluation Protocol

One of the challenges in benchmarking LLMs for mathematical reasoning is the inconsistency across evaluation metrics and protocols used in different work (Touvron et al., 2023; Luo et al., 2023; Yue et al., 2023).

MWPBENCH³ aims to evaluate the mathematical reasoning abilities of instruction tuned LLMs using a unified evaluation protocol. We employ zero-shot setting for evaluation and use the accuracy metric. The reason behind that is we believe fine-tuned LLMs should be able to answer questions directly without demonstrations, while in few-shot setting the final results may change with different set of demonstrations. For prompt template, we choose the Alpaca template (Taori et al., 2023) as default, which is the most widely used for instruction tuning (Taori et al., 2023; Luo et al., 2023; Yu et al., 2023). However, we support customized template just in case that LLMs are trained with a different instruction template (e.g., OpenAI ChatGPT template). For decoding, we choose greedy decoding to eliminate randomness in comparisons, selecting the top-1 completion as the solution. To further standardize the evaluation, we carefully implemented the answer extraction and verification processes (with high precision fuzzy match).

¹<https://docs.mathpix.com/#process-a-pdf>

²Copyright (c) by Michael J. Evans and Jeffrey S. Rosenthal. It may be copied and distributed without restriction, provided it is not altered, appropriate attribution is given and no money is charged.

³MWPBENCH is available at <https://github.com/microsoft/unilm/tree/master/mathscale>

Dataset	Level	Difficulty	Question Type	Language	#Train	#Test
GSM8K	Elementary	Easy	Word	En	7473	1319
MATH	Competition	ExHard	Word	En	7498	5000
TAL-SCQ	K12 Math	Medium	MC→Word	En	2638	1496
Math23k	Elementary	Easy	Word	Zh→En	1000	949
Ape210k	Elementary	Easy	Word	Zh→En	967	4874
GaokaoBench-Math	High School	Hard	MC→Word	Zh→En	0	508
AGIEval-Gaokao-Math	High School	Hard	MC→Word	Zh→En	0	404
AGIEval-SAT-Math	High School	Hard	MC→Word	En	0	102
CollegeMath	College	ExHard	Word	En	1281	2818
Total	–	–	–	–	20857	17470

Table 1. Statistics in MWPBENCH. In the “Question Type” column, “Word” stands for math word problem and “MC” stands for multiple-choice problem. In the “Difficulty” column, “ExHard” stands for extremely hard.

Topic	Book	License	#Train	#Test
Algebra	Beginning and Intermediate Algebra (Wallace, 2010)	CC BY 3.0	1171	1000
Precalculus	PRECALCULUS (Stitz & Zeager, 2013)	CC	80	500
Calculus	Calculus (Guichard, 2009)	CC BY-NC-SA	30	500
VectorCalculus	CORRAL’s VECTOR CALCULUS (Corral, 2008)	GFDL	0	110
Probability	Introduction to Probability (Grinstead & Snell, 2006)	GFDL	0	38
Probability	Probability and Statistics: The Science of Uncertainty (Evans & Rosenthal, 2004)	Custom ²	0	101
LinearAlgebra	Matrix Theory and LINEAR ALGEBRA (Selinger, 2018)	CC BY	0	123
LinearAlgebra	A First Course in LINEAR ALGEBRA (Kuttler & Farah, 2017)	CC BY	0	137
DifferentialEquations	ELEMENTARY DIFFERENTIAL EQUATIONS (Trench, 2001)	CC BY-NC-SA 3.0	0	309

Table 2. Details of permissively licensed books we use to construct the CollegeMath dataset.

3. MathScale: Scaling Instruction Tuning for Mathematical Reasoning

We present details of MathScale in this section. MathScale aims to generate large scale Mathematical Reasoning dataset by prompting ChatGPT and it contains four steps.

3.1. Concept Extraction

As shown in Figure 1, MathScale takes seed math questions as input and we use the training set of MWPBENCH (around 20K math questions). In the first step, we extract high level concepts (i.e., topics and knowledge points) from these seed questions with prompt engineering of GPT-3.5. We aim to extract meta information needed to solve a particular math question. We believe “topics” and “knowledge points” are important meta information for questions. A “topic” refers to the mathematical subject name or the topic name of math book chapter such as “Money and finance” and “Arithmetic operations”. While “knowledge points” refers to more fine grained math concepts (e.g., theorems, skills) in problem solving. Typical examples are “Definition and properties of dot product” or “Converting fractions to whole numbers”. We instruct GPT-3.5 to act as a Math teacher and extract 1 or 2 topics and 1 to 5 knowledge points from a given seed question (see the prompt template in Table 3).

To ensure the diversity of the extracted topics and knowl-

Act as a Math Teacher and analyze the provided question. Start by identifying 1 or 2 general topics that a student is being assessed on. Next, highlight 1 to 5 specific knowledge points that the question evaluates.

Provided question: {seed_question}

Analysis:

Table 3. Prompt template for Concept Extraction.

edge points, we use the training set of MWPBENCH, which includes questions from different sources. We also remove topics and knowledge points that appear only one time to reduce noise. In total, we extracted around 2K topics and 8K knowledge points. The above process mirrors the *concept compression* described in (Tall, 2013).

3.2. Concept Graph Construction

Concept Graph Given the topics and knowledge points extracted from the previous step, we move on to construct a concept graph C , whose nodes are the extracted topics $\mathbb{T} = \{t_1, t_2, \dots, t_{|\mathbb{T}|}\}$ and knowledge points (KPs) $\mathbb{K} = \{k_1, k_2, \dots, k_{|\mathbb{K}|}\}$. As shown in Figure 2, we have three types of edges in this graph (i.e., topic to topic edge, topic to KP edge and KP to KP edge), which results to three sub-graphs (topic graph, topic-KP graph, KP graph). When

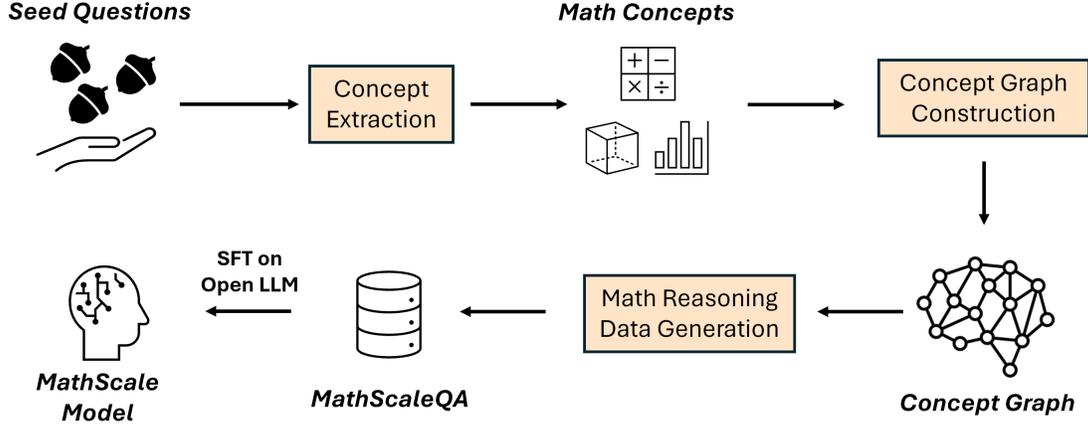


Figure 1. Overview of MathScale. MathScale starts from seed math questions and there are three steps in this pipeline (i.e., *concept extract*, *concept graph construction* and *mathematical reasoning data generation*). After these three steps, we obtain the MathScaleQA dataset, which is subsequently used to train open LLMs. Finally, we obtain MathScale models.

a topic (or KP) \mathbf{u} is co-occurred with another topic (or KP) \mathbf{v} , we build an edge between them and the edge weight is related to their co-occurrence statistics. Define co-occurrence as \mathbf{u} and \mathbf{v} have been extracted from the seed question.

Formally, let $E = \{(\mathbf{u}, \mathbf{v}) | f_{co}(\mathbf{u}, \mathbf{v}) > 0\}$ denote edges in C and $f_{co}(\mathbf{u}, \mathbf{v})$ is the edge weight between \mathbf{u} and \mathbf{v} . Intuitively, two KPs (or topics) are more likely to be reasonable composition when they have been frequently used to solve the same seed questions. Let $w_{\mathbf{u}\mathbf{v}}$ denote the raw co-occurrence count between node \mathbf{u} and node \mathbf{v} . The adjusted weight $f_{co}(\mathbf{u}, \mathbf{v})$ is defined as follows:

$$f_{co}(\mathbf{u}, \mathbf{v}) = \log(w_{\mathbf{u}\mathbf{v}} + \varepsilon) \quad (1)$$

where ε is a small constant introduced to maintain non-zero counts and prevent computational issues.

Concept Composition Given the graph C , we are ready to sample topics and KPs from it and the sampled topics and KPs are subsequently used to generate new math questions. We use a graph random walk algorithm to create concept compositions.

We start from a uniformly random sampling from the $|\mathbb{T}|$ topics we have extracted. Note that in implementation, we simply enumerate all extracted topics for multiple epochs.

In the second step, we do a random walk for one to two steps in the topic sub-graph to search for related topics. The probability distribution for the graph random walk is not uniform and defined as follows:

$$p_{\mathbf{u}\mathbf{v}} = \frac{\exp(f_{co}(\mathbf{u}, \mathbf{v}))}{\sum_{\mathbf{v}' \in \mathcal{N}(\mathbf{u})} \exp(f_{co}(\mathbf{u}, \mathbf{v}'))} \quad (2)$$

where $\mathcal{N}(\mathbf{u})$ denotes the set of nodes adjacent to \mathbf{u} in the **topic** sub-graph.

In the third step, we continue to randomly walk in the hybrid topic-KP graph for a single step with the probability distribution calculated as in Equation (2) on the **topic-KP** graph. So that we now have one sampled KP.

In the last step, we continue to expand to more KPs by randomly walking on the KP graph for zero to four steps again with the probability distribution computed as in Equation (2) on **KP** graph. We finally obtained a set of sampled topics $\hat{\mathbb{T}}$ and KPs $\hat{\mathbb{K}}$.

The whole process above is an analogy of the *connection forging* described in (Tall, 2013).

3.3. Mathematical Reasoning Data Generation

Act as a Math Teacher and create a new question and its solution based on the provided topics and knowledge points. Ensure that the created questions:

1. Adhere to the provided topics.
2. Necessitate the combined use of the associated knowledge points.

{few_shot_examples}

Topics:
{topics}

Knowledge Points:
{knowledge_points}

Structure your response as:
FORMAT INSTRUCTIONS OF THE NEW QA-PAIR ...

Table 4. Prompt template for Mathematical Reasoning Data Generation.

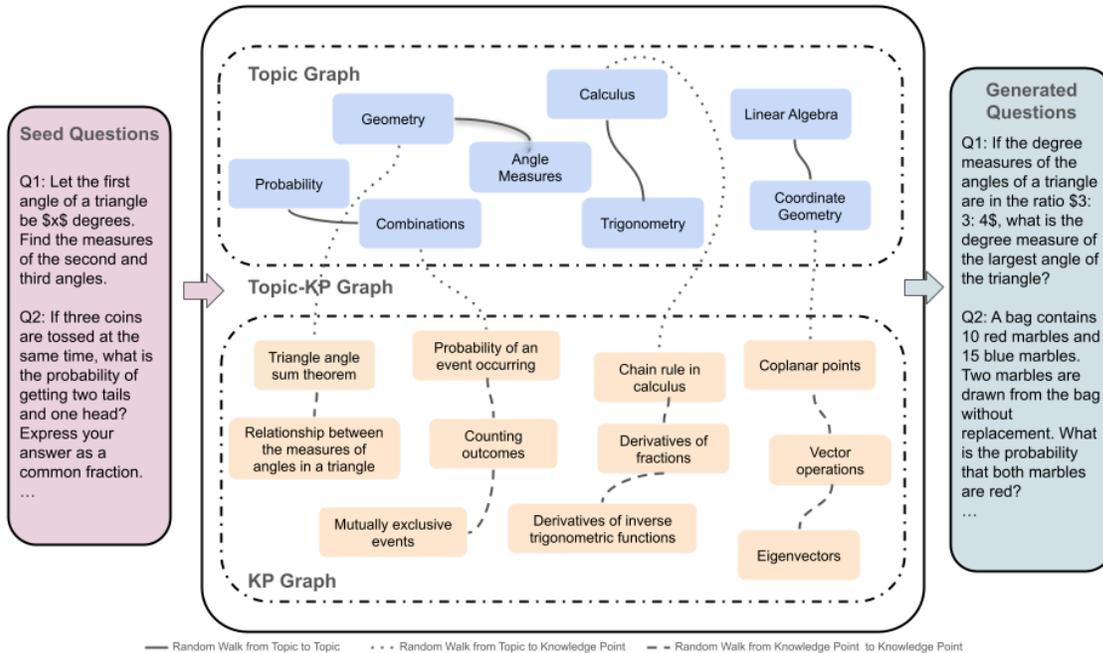


Figure 2. Running Examples of the concept graph construction process in the MathScale pipeline.

With the novel compositions of topics $\hat{\mathbb{T}}$ and KPs $\hat{\mathbb{K}}$ at hand, we query GPT-3.5 to generate corresponding question-answer pairs. Inspired by how math teachers design questions from existing exercises, we opt to include few-shot examples to guide GPT-3.5 in question formulation. These examples are chosen from the seed questions, based on the Jaccard distance of their knowledge points set. We ask GPT-3.5 to adhere to $\hat{\mathbb{T}}$ and encourage combine use of KPs $\hat{\mathbb{K}}$. We present the template for prompts in Table 4.

Furthermore, we apply a decontamination process, where all math questions in the test set of MWPBENCH are removed.

3.4. Validation

We observe that sometimes in the newly generated QA pairs, the solution is incorrect. We therefore also tried to add an additional validation process as follows. We first instruction GPT-4 to generate a reference solution for the question and then ask GPT-4 again to validate the GPT-4 solution against the solution generated in the previous step. We assume GPT-4 is more accurate than GPT-3.5. If GPT-4 believe the original solution is incorrect, we replace it with the new GPT-4 solution. Small scale experiments (Table 7) show the step does not improve the results. Perhaps because essentially we are trying to distill GPT-3.5 using open source LLMs. Although some solutions are incorrect, they are still help open source LLMs to learn the model distributions of GPT-3.5. Therefore, in our final pipeline, we remove this validation step.

4. Experiments

4.1. Implementation

Data Generation In concept extraction (Section 3.1), we use the MWPBENCH training set, comprising around 20K questions, as the seed questions for our MathScale pipeline and we employ GPT-3.5-Turbo-0613 for the extraction. In total, we obtain 2,018 topics and 8,892 knowledge points. We then construct graphs to establish relationships among these concepts (Section 3.2). The edge weight in the graph is smoothed using Equation (1) and we set $\varepsilon = 1e - 5$. In the concept composition process, treating the iteration through all topic nodes as one epoch, we repeat this process for approximately 1K epochs, resulting 2 million unique concept compositions. Then we instruct GPT-3.5-Turbo-0613 to create 2 million question-answer pairs with these compositions. We also decontaminate the generated datasets by excluding all math questions in the test set of MWPBENCH. To leverage the precious high quality math reasoning data, we additionally combine the generated data with the training set of MWPBENCH. We call the resulting dataset **MathScaleQA**. The validation step (Section 3.4) is excluded from the final pipeline, because we find that the validation step does not improve results (see details in Section 5.3). We provide concrete examples of concept extraction and MathScaleQA in Appendix A.4.

Model Training The questions in MathScaleQA are formatted using the Alpaca prompt (Taori et al., 2023) as follows.

Below is an instruction that describes a task. Write a response that appropriately completes the request.

```
### Instruction:
{question}
```

```
### Response:
```

Our training pipeline is adapted from the open-instruct (Wang et al., 2023) toolkit. We utilize the LLaMA-2 7B and 13B models (Touvron et al., 2023) as well as the Mistral 7B model (Jiang et al., 2023) as our backbone models. We use a batch size of 128 and train on the MathScaleQA dataset for 3 epochs using a learning rate of $2e-5$. We call the resulting models *MathScale-7B*, *MathScale-13B* and *MathScale-Mistral-7B*. We leave exploration of the LLaMA-2 70B model in future work.

4.2. Models in Comparison

Close-Source Models We include the most capable GPT models developed by OpenAI, which are the light-weighted GPT-3.5-Turbo-0613 and the powerful GPT-4-0314. These models are known to be good at mathematical reasoning and serves as the upper bounds.

Open-Source Models: We also compare our model against open-source math models. Specially, we compare with WizardMath (Luo et al., 2023), GAIR-Abel (Chern et al., 2023), MetaMath (Yu et al., 2023), and MAMmoTH (Yue et al., 2023). WizardMath (Luo et al., 2023) is based on evol-instruct (Xu et al., 2023) and reinforcement learning. MetaMath (Yu et al., 2023) is trained on a dataset by augmenting GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) using answer or question side paraphrasing. The dataset used to train MAMmoTH (Yue et al., 2023) comprises a collection of 13 existing math datasets with GPT-4 CoT (Wei et al., 2022) and/or PoT (Gao et al., 2023; Chen et al., 2022) annotations. We evaluate all models using CoT natural language style math solutions. We noticed that some of the models (e.g., GPT-4 and MAMmoTH) can produce code solution of math problems in addition to natural language solutions. For fair comparison, we refrain from comparing using code-interpreter style solutions, because all models above can produce code-interpreter style solutions if the solutions in their training data are replace by GPT annotated code solutions. Also note that WizardMath v1.1 is a Mistral based math model and we do not know how its training data are constructed (the authors did not release any detail of the training data of WizardMath v1.1). We evaluate all models on MWPBENCH, which contains 9 datasets on mathematical reasoning. We report accuracies of the 9 datasets as well as their micro-average and macro-average. We prompt all models using the Alpaca template (see Section 4.1). (Luo

et al., 2023) recommended an improved prompt for during inference (i.e., adding Let’s think step by step after the standard Alpaca template). However, we observe mixed results on MWPBENCH for some models in comparison. For example, we observe improved results on GSM8K, but decreased results on MATH. We therefore do not use this optimization for all models in comparison.

4.3. Main Results

As shown in Table 5, MathScale obtains best micro average and macro average scores on MWPBENCH compared to other models based on LLaMA-2 7B, LLaMA-2 13B or Mistral 7B. Specifically, On average, *MathScale-7B* achieves a 35.2% (micro) and 38.2% (macro) accuracy across MWPBENCH, surpassing its best counterparts of equivalent size by 42.8% and 43.6%, respectively. The trends are similar for *MathScale-13B* and *MathScale-Mistral*. This also confirms the effectiveness of our MathScaleQA dataset regardless of the backbone model. Note that in GaokaoBench-Math, AGIEval-Gaokao-MATH, and AGIEval-SAT-MATH, there is no training set. Even on these out-of-domain test sets, *MathScale-7B* wildly outperforms other open-source models in comparison. When compared to frontier LLMs, MathScale-Mistral demonstrates performance parity in both micro and macro averages relative to GPT-3.5-Turbo (see the first block in Table 5). We further compare with WizardMath and MetaMath under the same training size (see Figure 3 and detailed results in Appendix A.6), and that *MathScale-7B* consistently achieves better results on most datasets as well as both macro and micro accuracies. Results of different subsets on MATH and CollegeMath are in Appendix A.7.

5. Analysis and Discussions

5.1. Scaling Property of MathScale

As described in Section 3, given a fixed set of math concepts, iterating over concept graphs allows us to generate different compositions of mathematical concepts, thereby synthesizing large amount of new math data. We use LLaMA-2 7B as our base model to study the scaling property of MathScale. When scaling the size of the MathScaleQA dataset, we observe a nearly logarithmic growth in the performance of the MathScale-7b model across all datasets within MWPBENCH, as depicted in Figure 3. We draw the scaling curve up to two million examples (size of the full MathScaleQA). We also compare MathScale against WizardMath and MetaMath at their respective training sizes. MathScale outperforms both models across all datasets (except for GSM8K) when using an equivalent amount of training data. Given the scaling curves in Figure 3, we anticipate that the performance of MathScale may continue to improve with even more synthetic training examples. Due to resource constraints, we

Models	GSM8K	MATH	College Math	TAL	Math23k	Ape210k	Gaokao Bench Math	AGIE Gaokao Math	AGIE SAT Math	Micro Average	Macro Average
<i>Closed-source Models</i>											
GPT-4	92.9	51.8	24.4	51.8	76.5	61.5	35.4	28.2	68.6	52.0	54.5
GPT-3.5-Turbo	74.1	37.8	21.6	42.9	62.5	44.0	23.2	15.3	55.8	39.9	41.9
<i>Models based on LLaMA-2 13B</i>											
LLaMA-2 13B	7.1	3.5	1.2	6.3	9.5	7.9	0.7	0.4	6.8	4.5	4.8
WizardMath	62.0	14.3	7.8	18.7	38.3	25.2	8.2	3.4	29.4	20.4	23.0
MAmmoTH-CoT	56.5	12.6	6.5	17.3	39.5	28.1	5.9	4.9	20.5	19.3	21.3
GAIR-Abel	66.4	16.6	7.9	21.1	42.2	27.8	7.0	4.9	30.3	22.5	24.9
MetaMath	70.8	22.8	10.1	25.4	48.6	31.6	9.6	5.6	38.2	27.0	29.1
MathScale 13B	71.3	33.8	20.4	38.1	61.1	43.7	20.0	12.3	55.8	37.2	39.6
<i>Models based on LLaMA-2 7B</i>											
LLaMA-2 7B	4.5	4.2	2.3	7.6	6.8	7.3	2.1	2.9	2.9	4.6	4.5
WizardMath	52.8	10.3	6.8	14.0	32.5	19.2	5.9	6.1	22.5	16.3	18.9
MAmmoTH-CoT	50.0	9.5	6.2	13.3	34.6	21.4	3.9	2.7	19.6	15.8	17.9
GAIR-Abel	57.6	12.7	6.6	18.3	35.4	24.5	4.3	4.4	23.5	18.7	20.8
MetaMath	66.2	20.6	9.4	22.5	44.0	29.9	5.9	5.1	36.2	24.7	26.6
MathScale 7B	66.3	31.1	20.9	35.2	59.0	41.8	19.6	12.6	57.8	35.2	38.2
<i>Models based on Mistral 7B</i>											
Mistral 7B	15.5	10.1	7.5	17.9	18.5	15.5	6.2	5.9	22.5	12.0	13.2
WizardMath v1.1	78.1	32.8	16.0	34.4	58.3	41.4	16.1	9.6	55.8	35.5	38.0
MetaMath Mistral	77.4	28.4	15.7	31.4	55.1	38.1	15.3	10.1	50.9	32.9	35.8
MathScale Mistral	74.8	35.2	21.8	39.9	64.4	46.0	21.4	14.3	57.8	39.1	41.7

Table 5. Performance metrics on MWPBENCH. All evaluations were conducted utilizing the driver provided by MWPBENCH, ensuring a consistent and fair comparison. Within each section, the highest results are highlighted in bold font. “AGIE” stands for AGIEval.

leave the training set scaling beyond two million examples to future work.

5.2. Ablation on Concept Extraction

In the concept extraction process (Section 3.1), we use all the 20K seed questions. We attempt to answer the following two questions. 1) *Does the number of seed questions matter?* 2) *Does the number of extracted concepts matter?* We control the size of resulting training examples to 25K for fast experimentation. In all experiments, we use the LLaMA-2 7B model as our backbone model.

Number of Seed Questions To assess the influence of seed questions, we firstly randomly remove 50% of the seed questions from the MWPBENCH training set (i.e., we use only 10K seed questions). The results are shown in Table 6. We observe the macro average on MWPBENCH drops by 2.9%. Further, when we limite the data source of seed questions exclusively to the training sets of GSM8K and MATH, there is a performance decrease of 3.5%. These results above indicate that incorporating of a larger and more diverse set of seed questions is beneficial.

Number of Math Concepts Additionally, we examine the impact of extracted math concepts. As shown in Table 6, by removing half of the topics or knowledge points, we observe a notable decrease in the macro average on the MWPBENCH. Particularly, removing knowledge points lead to a greater decrease in performance (i.e., -8.6% with 50% knowledge points v.s. -2.3% with 50% of topics). This highlights the essential role that knowledge points play in enhancing the effectiveness of MathScale.

Methods	Macro Average	Relative Change
MathScale	14.5	-
Remove 50% Seed Questions	14.0	-2.9%
Restrict Data Source to GSM8K and MATH only	13.9	-3.5%
Remove 50% Topics	14.1	-2.3%
Remove 50% Knowledge Points	13.2	-8.6%

Table 6. Ablation studies of concept extraction with a control training size of 25K on MWPBENCH.

5.3. On Validating Generated Data

The generated QA pairs in MathScaleQA might be incorrect. Therefore, we introduce a separate validation step in Sec-

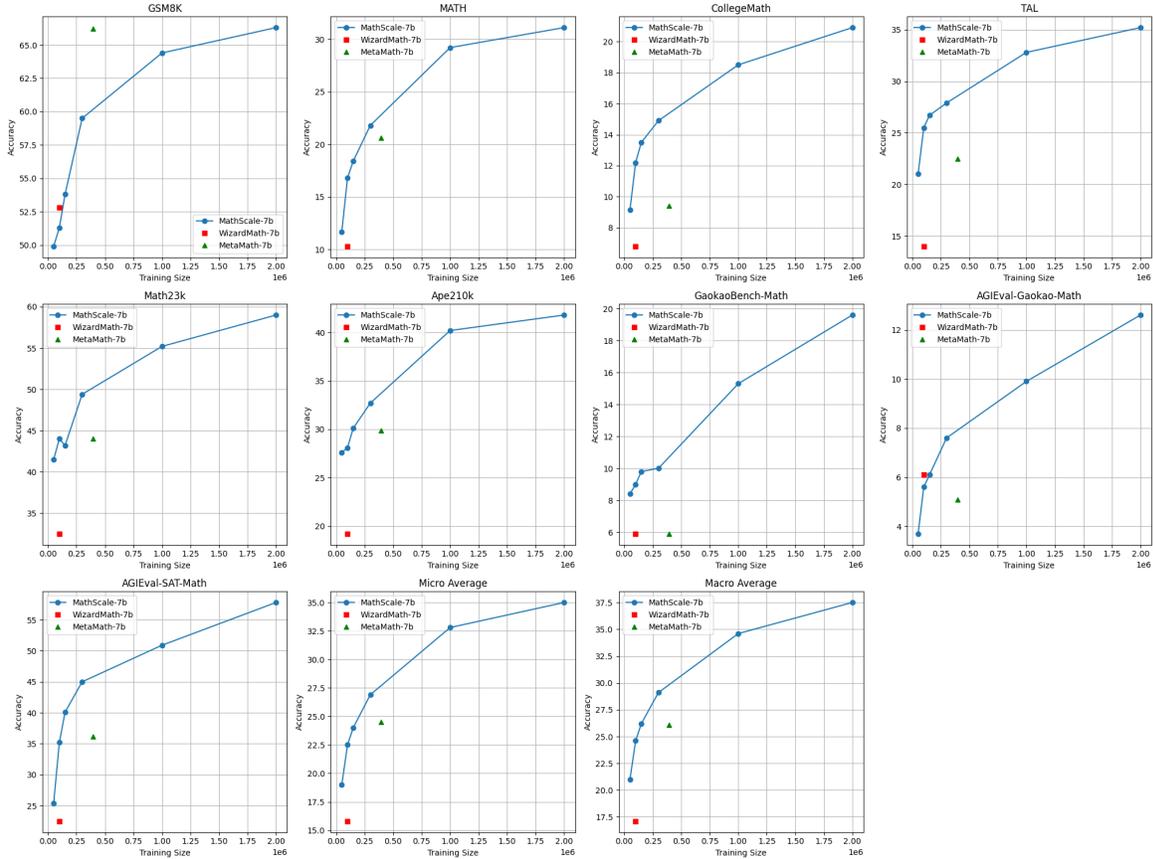


Figure 3. Performance on MWPBENCH using different sizes of training dataset in MathScaleQA.

tion 3.4. In this section, we design controlled experiment on 5K generated data from MathScaleQA and again using LLaMA-2 7B as our base model.

GPT-4 vs. GPT-3.5 Accuracy We manually annotate 100 randomly chosen generated data points and generate answers with GPT-3.5-Turbo and GPT-4. GPT-4 demonstrate an impressive accuracy of 87%, significantly outperforming the accuracy of 69% by GPT-3.5-Turbo. Therefore, we used GPT-4 to generate reference solutions and validate our synthetic solutions, replacing any incorrect solutions with the GPT-4 reference solutions.

Results Within the 5K examples, 26% of the solutions are identified as incorrect by GPT-4 and are replaced. We have another two settings with either all GPT-3.5 solutions and GPT-4 solutions. The results are shown in Table 7 and we observe that using original 3.5-Turbo solutions lead to a similar results as using the validation step.

This observation is counter-intuitive. Maybe because training on synthetic data generated from GPT-3.5 is essential distillation. Even if some solutions are incorrect, they may still help to the open-source LLMs (e.g., LLaMA-2 or Mis-

tral) to mimic the distributions of GPT-3.5. We also notice that in neural machine translation distillation, the step of validating incorrect translations is also ignored (Kim & Rush, 2016). Considering the high cost of GPT-4 and similar results obtained, we opt to omit the validation and correction step from the final MathScale pipeline⁴.

5.4. Performance on a Fresh Math Dataset

While MathScaleQA generated by GPT-3.5 is rigorously decontaminated to prevent overlap with the MWPBENCH test set, there may still be small chance that some of the test sets have been leaked to GPT-3.5-Turbo or contained in the training data of LLaMA-2. Because GPT-3.5-Turbo uses human annotated queries submitted by users through their APIs⁵. These queries may include test sets such as GSM8K. The training set of LLaMA-2 is not released and we are not sure if some examples in test sets of MWPBENCH are included or not.

⁴We are not trying to draw the conclusion that if we replace GPT-3.5 with GPT-4 in all steps of MathScale (see more details in Appendix A.5).

⁵<https://openai.com/research/instruction-following>

Methods	Micro Average	Macro Average
100% GPT-3.5 Solutions	10.6	11.5
74% GPT-3.5 Solutions and 26% GPT-4 Corrected Solutions	10.2	11.1
100% GPT-4 Solutions	9.8	10.9

Table 7. Ablation studies of validation step with a control training size of 5K on MWPBENCH.

To address this issue, we manually curate a new dataset comprising the latest 30 math questions from latest Gaokao Math exam, held in June for China National Higher Education Entrance Examination. We term this dataset, *Fresh-GaokaoMath-2023*, which we believe Fresh-GaokaoMath-2023 is not likely to be included in the training data of LLaMA-2 or GPT-3.5-Turbo. Because LLaMA-2 and GPT-3.5-Turbo are released before Fresh-GaokaoMath-2023 is created.

We compare our LLaMA-2 7B based model MathScale-7B against two other LLaMA-2 7B based models (i.e., WizardMath-7B and MetaMath-7B) as well as GPT-3.5-Turbo and GPT-4. Results are in Table 8. MathScale consistently surpasses WizardMath and MetaMath, which aligns with the main results shown in Table 5. It demonstrates the robustness and adaptability of MathScale in handling fresh math questions.

Model	Fresh-GaokaoMath-2023
GPT-4	43.3
GPT-3.5-Turbo	40.0
WizardMath-7B	13.3
MetaMath-7B	16.6
MathScale-7B	30.0

Table 8. Performance metrics on Fresh-GaokaoMath-2023.

6. Related Work

ChatGPT-based Instruction Tuning A pivotal aspect driving advancements in math instruction tuning is the use of ChatGPT for data synthesis. For instance, WizardMath (Luo et al., 2023) introduced reinforced evol-instruct which integrates five operations: adding constraints, deepening, concretizing, increasing reasoning steps, and complicating input, thereby facilitating comprehensive evolution. Similarly, MetaMath (Yu et al., 2023) employs a bootstrapping strategy for questions, incorporating answer augmentation, rephrasing, self-verification, and FOFAR. While these methods are effective, the breath space is inherently confined to manually designed operations. Our approach seeks to enable ChatGPT to emulate cognitive processes in human mathematical learning, thus overcoming the limitations faced by previous methodologies.

Tool-Integration Instruction Tuning Recent studies have also explored integrating tools into ChatGPT-based instruction tuning for mathematics. ToRA (Gou et al., 2023) com-

bines natural language reasoning with program-based tool usage to synthesize trajectory data. Each trajectory iteratively concatenates reasoning, programming, and program outputs until the final answer is reached. Our current focus is solely on natural language reasoning. While tool integration within the MathScale pipeline is an intriguing prospect, we reserve its exploration for future research.

7. Conclusions

We propose *MathScale*, a simple and scalable method to create high-quality mathematical reasoning data using frontier LLMs. We also construct MWPBENCH, a comprehensive benchmark of Math Word Problems covering K-12, college, and competition level math problems. Evaluated on MWPBENCH, MathScale-7B achieves state-of-the-art performance across all datasets, surpassing its best peers of equivalent size by 42.8% in micro average accuracy and 43.6% in macro average accuracy, respectively.

Acknowledgements

This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing.

Impact Statement

This paper seeks to advance mathematical reasoning by introducing a scalable method for generating high-quality synthetic data with large language models, along with new evaluation benchmarks to foster consistent and fair model comparisons in academia. While our efforts center on assessing mathematical capabilities, it’s crucial to note that the models may exhibit biases not examined in our study. Addressing these biases and ensuring the models’ alignment with societal values is essential, highlighting the need for comprehensive evaluations that encompass both technical performance and ethical considerations.

References

- Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.
- Chern, E., Zou, H., Li, X., Hu, J., Feng, K., Li, J., and Liu, P. Generative ai for math: Abel. <https://github.com/GAIR-NLP/abel>, 2023.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Corral, M. CORRAL’S VECTOR CALCULUS. 2008.
- Evans, M. J. and Rosenthal, J. S. *Probability and statistics: The science of uncertainty*. Macmillan, 2004.
- Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. PAL: Program-aided language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/gao23f.html>.
- Gou, Z., Shao, Z., Gong, Y., Yang, Y., Huang, M., Duan, N., Chen, W., et al. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.
- Grinstead, C. M. and Snell, J. L. *Grinstead and Snell’s introduction to probability*. Chance Project, 2006.
- Guichard, D. *Calculus*. 2009.
- Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., et al. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021a.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021b.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Kim, Y. and Rush, A. M. Sequence-level knowledge distillation. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL <https://aclanthology.org/D16-1139>.
- Kuttler, K. and Farah, I. *A First Course in Linear Algebra, 2017A version (Lyryx)*. Lyryx, 2017.
- Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., and Zhang, D. Wizard-math: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.
- Mishra, S., Finlayson, M., Lu, P., Tang, L., Welleck, S., Baral, C., Rajpurohit, T., Tafjord, O., Sabharwal, A., Clark, P., et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022.
- Selinger, P. Matrix theory and linear algebra, 2018. URL <https://www.mathstat.dal.ca/~selinger/linear-algebra/>. An introduction to linear algebra for first or second year university students. Licensed under Creative Commons CC BY 4.0 License. Last updated on October 26, 2018.
- Stitz, C. and Zeager, J. *Precalculus*. Stitz Zeager Open Source Mathematics, 2013.
- TAL. Tal-scq5k, 2023. URL <https://github.com/math-eval/TAL-SCQ5K>. GitHub repository.
- Tall, D. *How humans learn to think mathematically: Exploring the three worlds of mathematics*. Cambridge University Press, 2013.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

- Trench, W. F. *Elementary Differential Equations*. Brooks/Cole Thomson Learning, San Antonio, Texas, USA, 2001.
 URL http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_DIFF_EQNS_I.PDF.
 Free Edition 1.01 (December 2013).
- Wallace, T. *Beginning and intermediate algebra*. 2010.
- Wang, Y., Liu, X., and Shi, S. Deep neural solver for math word problems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 845–854, 2017.
- Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.
- Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.
- Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.
- Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X. Evaluating the performance of large language models on gaokao benchmark. 2023.
- Zhao, W., Shang, M., Liu, Y., Wang, L., and Liu, J. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*, 2020.
- Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

A. Appendix

A.1. MWPBENCH: Transform Non-Word Problems into Word Problems

For datasets like TAL-SCQ (TAL, 2023), GaokaoBench-Math (Zhang et al., 2023), and AGIEval (Zhong et al., 2023), the problems are presented in a multiple-choice format. To eliminate the influence of the problem type and concentrate on the intrinsic ability of LLMs to address mathematical problems, we converted these non-word problems into word problems.

A.1.1. FILTERING QUESTIONS

Initially, we identified and filtered out questions that rely heavily on the multiple-choice format. This filtering was done using specific keywords and phrases that are indicative of multiple-choice questions.

```

1 def is_bad_question(question):
2     question = question.lower()
3
4     keywords = [
5         "?",
6         "which_of_the_following",
7         "which_one",
8         "which_is",
9         "the_following",
10        "which_statement"
11    ]
12
13    for keyword in keywords:
14        if keyword in question:
15            print(f"Filtered_question:_{question}")
16            return True
17    return False

```

Listing 1. Filtering questions

A.1.2. CREATING QUESTION-ANSWER PAIRS

After filtering out the aforementioned questions, the remaining questions were paired with their corresponding correct answer choices. This transformation resulted in a format where each problem is presented as a word problem followed by its solution.

A.2. MWPBENCH: Translation of Non-English Problems to English

For several datasets, namely Math23k (Wang et al., 2017), Ape210k (Zhao et al., 2020), GaokaoBench-Math (Zhang et al., 2023), and AGIEval-Gaokao (Zhong et al., 2023), the problems are originally presented in Chinese. To ensure uniformity and mitigate the effects of multilingual representations, we translated these Chinese problems into English. The translation was facilitated by the GPT-3.5-Turbo API. Due to parsing errors encountered during the post-processing, a few examples were excluded. The prompt template employed for the translation request is provided below:

```

I want you to act as a Math Translator. Your task is to translate Chinese math questions into English math
questions.
Make sure to keep the original question numbers.
Make sure to keep the math formula in Latex format.
The translations should be clear, accurate, and easily understandable for students who are native English speakers.

# Chinese Math Questions #:
<insert chinese questions>

# English Math Questions #:

```

A.3. CollegeMath: Extraction from textbooks

To construct the CollegeMath dataset, we made use of the GPT-3.5-Turbo API to parse and extract questions and answers from raw, segmented LaTeX exercises and their corresponding solutions.

A.3.1. EXTRACTING QUESTIONS FROM EXERCISES

The primary goal was to convert raw, potentially unstructured questions from math textbooks into well-formulated LaTeX-formatted questions. Below is the prompt template we utilized for this extraction process:

```
I want you to act as a Math Parser. Your task is to convert raw messy questions from a math textbook into well-structured LaTeX-formatted questions.
```

```
Please ensure to retain the original question numbers.
```

```
If needed, prepend the original instructions to the parsed questions to make them more comprehensible.
```

```
If needed, skip the broken questions.
```

```
<insert demo>
```

```
#Raw Questions#:
```

```
'''
```

```
<insert a chapter of practice>
```

```
'''
```

```
#Well-structured LaTeX-formatted Questions#:
```

A.3.2. EXTRACTING ANSWERS FROM SOLUTIONS

Similarly, for answers, our aim was to transform raw, messy answers from textbooks into clear, LaTeX-formatted answers. Here's the template for this task:

```
I want you to act as a Math Parser. Your task is to convert raw messy answers from a math textbook into well-structured LaTeX-formatted answers.
```

```
Please ensure to retain the original answer numbers.
```

```
If needed, skip the broken answers.
```

```
<insert demo>
```

```
#Raw Answers#:
```

```
'''
```

```
<insert a chapter of answer>
```

```
'''
```

```
#Well-structured LaTeX-formatted Answers#:
```

By employing the aforementioned prompt templates, we were able to extract a comprehensive set of questions and answers, thereby forming the foundation of the CollegeMath dataset.

A.4. MathScale: Concrete Examples

A.4.1. EXTRACTED TOPICS

A set of 30 topics, randomly chosen, is listed below to illustrate the variety:

```
"Arithmetic operations" "Word problem solving" "Mathematics" "Money and finance" "Problem-solving strategies"  
"Arithmetic" "Multiplication" "Proportions" "Basic arithmetic operations" "Conversion of units" "Measurement and weight"  
"Multiplication and addition" "Budgeting" "Basic arithmetic" "Wages and overtime" "Calculating earnings"
```

"Arithmetic Sequences" "Exponential Growth" "Financial calculations" "Problem solving" "Algebraic expressions" "Economics" "Time" "Business and finance" "Ratio and proportion" "Problem-solving" "Time calculations" "Addition" "Distance" "Speed"

A.4.2. EXTRACTED KNOWLEDGE POINTS

Similarly, we provide a list of 30 knowledge points, chosen at random, to demonstrate the depth and breadth of content:

"Random selection of marbles" "Definition and properties of dot product" "Manipulation of complex numbers" "Calculation of time required to complete a task" "How to apply the concept of a seven-day cycle" "Distinct numbers" "Expectation of a function of a random variable" "Ability to calculate total time" "Combinations of numbers" "Calculation of weekly income" "Relative motion" "Understanding the relationship between centimeters and kilometers" "Diagonalizing a matrix" "Proportional relationships between two quantities" "Ergodic Markov chain" "Addition of values" "Counting the number of cars" "Converting fractions to whole numbers" "Identifying relationships between different variables" "Ability to set up and solve a proportion equation" "Addition and subtraction of matrices" "Using logarithms to solve exponential equations" "Probability of rolling a specific number on a six-sided die" "Divisibility of polynomials" "Application of multiplication to calculate total revenue" "Identifying the highest and lowest scores" "Ability to calculate percentages." "Geometric interpretation of dot product" "Dividing complex numbers" "Understanding weight units"

A.4.3. A TRAINING EXAMPLE OF MATHSCALEQA

We provide a concrete training example from MathScaleQA:

-----gpt-3.5 generated question wrapped in alpaca-----
Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
Given the matrix $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 2 \\ 3 & 1 & 4 \end{pmatrix}$, find the rank of \mathbf{A} .

Response:
-----gpt-3.5 generated answer-----
To find the rank of a matrix, we need to find the largest number of linearly independent rows (or columns). We can use the row echelon form to help us determine this.

First, we perform the following row operations to obtain the row echelon form:
1. Subtract 2 times the first row from the second row.
2. Subtract 3 times the first row from the third row.

We obtain the matrix:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -4 \\ 0 & -5 & -5 \end{pmatrix}$$

Next, we perform the following row operations: 1. Add 5 times the second row to the third row.

We obtain the matrix:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -4 \\ 0 & 0 & -15 \end{pmatrix}$$

Now, our matrix is in row echelon form, and we can see that there are 3 non-zero rows, which means there are 3 linearly independent rows. Therefore, the rank of \mathbf{A} is 3.

The answer is 3.

A.5. Why Choose GPT-3.5?

We are not trying to draw the conclusion that if we replace GPT-3.5 with GPT-4 in all steps of MathScale, we will get worse results. On the contrary, we strongly believe that upgrading GPT-3.5 to GPT-4 for the whole MathScale pipeline, the results are highly likely to be better (with potentially better topic/KP extraction and better question generation). We use GPT-3.5 in

our pipeline because 1) it is cheaper and therefore friendly to scaling experiments; 2) it is easier to compare with previous work (e.g., MetaMath, WizardMath), which are also based on GPT-3.5.

Why GPT-4 based solutions lead to slightly worse results?

- the capability gap between LLaMA-2-7b and GPT-4 is larger than that between LLaMA-2-7b and GPT-3.5 (see Sec 5.3)
- we observed that the GPT-4 based solutions are 15% longer than GPT-3.5 based solutions on average, which might be harder to learn.

A.6. Comparison under the Same Training Size

Models	Train Size	GSM8K	MATH	College Math	TAL	Math23k	Ape210k	Gaokao Bench Math	AGIE Gaokao Math	AGIE SAT Math	Micro Average	Macro Average
WizardMath-7B	96.5K	52.8	10.3	6.8	14.0	32.5	19.2	5.9	6.1	22.5	16.3	18.9
MathScale-7b	96.5K	51.3	16.8	12.2	25.5	44.0	28.1	9.0	5.6	35.2	22.7	25.2
MetaMath-7b	395K	66.2	20.6	9.4	22.5	44.0	29.9	5.9	5.1	36.2	24.7	26.6
MathScale-7b	300K	59.5	21.8	14.9	27.9	49.4	32.7	10.0	7.6	45.0	27.1	29.8

Table 9. Comparison with WizardMath and MetaMath under the Same Training Size on MWPBENCH.

We show our final results with 2M examples in Table 5 to demonstrate how far our method can go by scaling up. For comparison with other models with the same number of examples, please refer to Figure 3. The last two figures show our method always obtain a better results compared to MetaMath and WizardMath on macro/micro average precisions with a comparable sized training examples. The detailed results are shown the Table 9. Note that we do not have results of using 395K examples for MathScale-7b. Therefore, we use the results of 300k instead to compare with MetaMath-7b. Even if we used fewer training examples, we still obtain better results on 8 out of 9 datasets.

A.7. Evaluation on Individual Topics

We examine the subset performances on MATH and CollegeMath, as shown in Table 10 and Table 11, respectively. It is evident that MathScale consistently delivers exceptional results across diverse topics. For CollegeMath, despite the MWPBENCH training set’s seed questions only encompassing algebra, precalculus, and calculus, MathScale demonstrates robust performance in OOD’s test sets including vector calculus, probability, and linear algebra. However, an area of challenge is differential equations, where all models show limited success.

We further provide an example of MathScale Mistral answering a question from MATH test set:

```
-----a MATH test question wrapped in alpaca-----
Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
Suppose that I have 6 different books, 2 of which are math books. In how many ways can I stack my 6 books on a shelf if I do not want the math books to be next to each other?

Response:
-----MathScale-Mistral generated answer-----
Let’s first find the total number of ways to stack the 6 books on the shelf, which is simply 6!. Now, let’s count the number of ways in which the math books are next to each other. We can treat the two math books as a single block, which can be stacked in 5! ways. However, there are 2! ways to arrange the math books within the block, so the total number of ways in which the math books are next to each other is 2! · 5!. Therefore, the number of ways in which the math books are not next to each other is 6! - 2! · 5! = 480.

The answer is 480.
```

MathScale: Scaling Instruction Tuning for Mathematical Reasoning

Model	MATH						
	Prealgebra	Algebra	Intermediate Algebra	Precalculus	Probability	Geometry	Number Theory
<i>closed-source models</i>							
GPT-4	75.2	71.3	25.3	30.4	52.5	41.7	45.7
GPT-3.5	59.3	55.5	17.3	20.1	30.1	29.8	30.3
<i>open-source models fine-tuned on LLaMA-2 13B</i>							
WizardMath	23.6	21.4	7.5	7.1	10.9	12.3	6.8
MAmmoTH-CoT	21.4	17.2	6.9	7.8	11.8	8.7	6.2
GAIR-Abel	28.3	23.3	8.1	9.1	13.0	15.0	9.4
MetaMath	39.3	32.1	11.9	10.2	18.5	17.7	15.3
MathScale	52.9	53.4	13.6	17.3	24.6	25.6	25.7
<i>open-source models fine-tuned on LLaMA-2 7B</i>							
WizardMath	16.5	15.2	6.3	5.8	6.7	8.5	5.9
MAmmoTH-CoT	15.1	12.5	6.5	4.3	9.9	7.3	6.1
GAIR-Abel	21.4	17.6	7.7	6.9	10.1	9.8	7.4
MetaMath	34.0	29.6	8.7	9.8	17.5	15.4	17.5
MathScale	48.9	49.3	12.4	15.2	23.2	23.3	23.8
<i>open-source models fine-tuned on Mistral 7B</i>							
WizardMath v1.1	51.4	50.7	13.9	19.9	25.5	24.4	22.4
MetaMath Mistral	47.1	41.4	13.2	12.6	23.4	23.7	19.8
MathScale Mistral	55.9	52.8	14.6	18.6	28.9	26.5	27.5

Table 10. Performance metrics across various topics on MATH. Within each section, the highest performing results are highlighted in bold font.

Model	CollegeMath						
	Algebra	Precalculus	Calculus	Vector Calculus	Probability	Linear Algebra	Differential Equation
<i>closed-source models</i>							
GPT-4	41.1	21.2	20.6	29.0	11.5	6.5	1.2
GPT-3.5	37.7	16.6	17.8	32.7	10.0	3.0	1.2
<i>open-source models fine-tuned on LLaMA-2 13B</i>							
WizardMath	12.0	7.4	8.2	14.5	2.8	0.3	0.3
MAmmoTH-CoT	11.2	4.2	7.0	8.1	2.8	1.5	0.0
GAIR-Abel	15.3	6.0	5.0	3.6	2.1	1.9	1.6
MetaMath	19.4	9.8	5.6	8.1	1.4	1.1	0.3
MathScale	35.0	17.8	15.8	24.5	7.9	5.0	1.9
<i>open-source models fine-tuned on LLaMA-2 7B</i>							
WizardMath	9.7	5.2	10.2	11.8	1.4	1.1	0.3
MAmmoTH-CoT	9.5	4.8	7.0	10.0	2.1	3.4	0.0
GAIR-Abel	12.0	4.2	5.2	6.3	3.5	1.5	1.6
MetaMath	19.1	6.8	4.4	5.4	2.8	2.6	0.3
MathScale	34.2	19.6	18.8	27.2	7.9	5.0	0.6
<i>open-source models fine-tuned on Mistral 7B</i>							
WizardMath v1.1	29.3	14.0	11.4	16.3	5.0	2.3	0.0
MetaMath Mistral	28.1	12.2	11.2	21.8	7.1	3.8	0.6
MathScale Mistral	37.1	18.0	19.4	27.2	8.6	3.8	1.6

Table 11. Performance metrics across various topics on CollegeMath. Within each section, the highest performing results are highlighted in bold font.