

MathScale: Scaling Instruction Tuning for Mathematical Reasoning

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

The Chinese University of Hong Kong, Shenzhen¹

Microsoft Research Asia²

Shenzhen Research Institute of Big Data³

Instruction Tuning for Mathematical Reasoning

(1) GSM8K & MATH

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

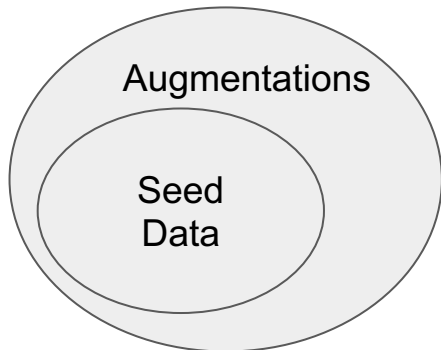
The most popular datasets, each only having 7.5K training data.

(2) Augmentations

- **Augment questions:** increased complexity
- **Augment answers:** diverse reasoning paths

[Luo, Haipeng, et al.\(2023\)](#) [Yu, Longhui, et al. \(2023\)](#)

Dependency on Seed Data



Augmentation on Questions

Question: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

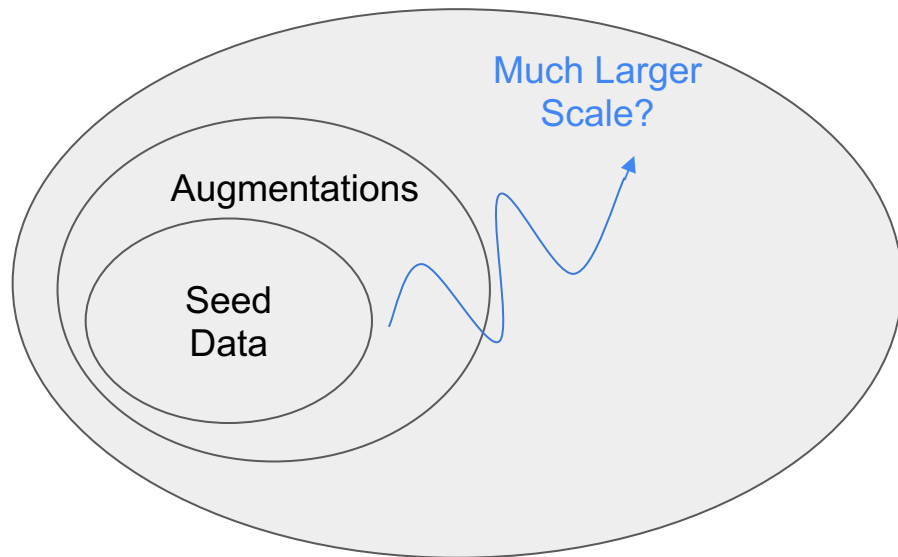
Rephrase the above question: What is the amount of money that Olivia has left after purchasing five bagels for \$3 each, if she initially had \$23?

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

Rephrase the above question: After losing 23 golf balls on Tuesday and an additional 2 on Wednesday, how many golf balls does Michael have left if he initially had 58 golf balls?

[Yu, Longhui, et al. \(2023\)](#)

Dependency on Seed Data



Augmentation on Questions

Question: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

Rephrase the above question: What is the amount of money that Olivia has left after purchasing five bagels for \$3 each, if she initially had \$23?

Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

Rephrase the above question: After losing 23 golf balls on Tuesday and an additional 2 on Wednesday, how many golf balls does Michael have left if he initially had 58 golf balls?

[Yu, Longhui, et al. \(2023\)](#)

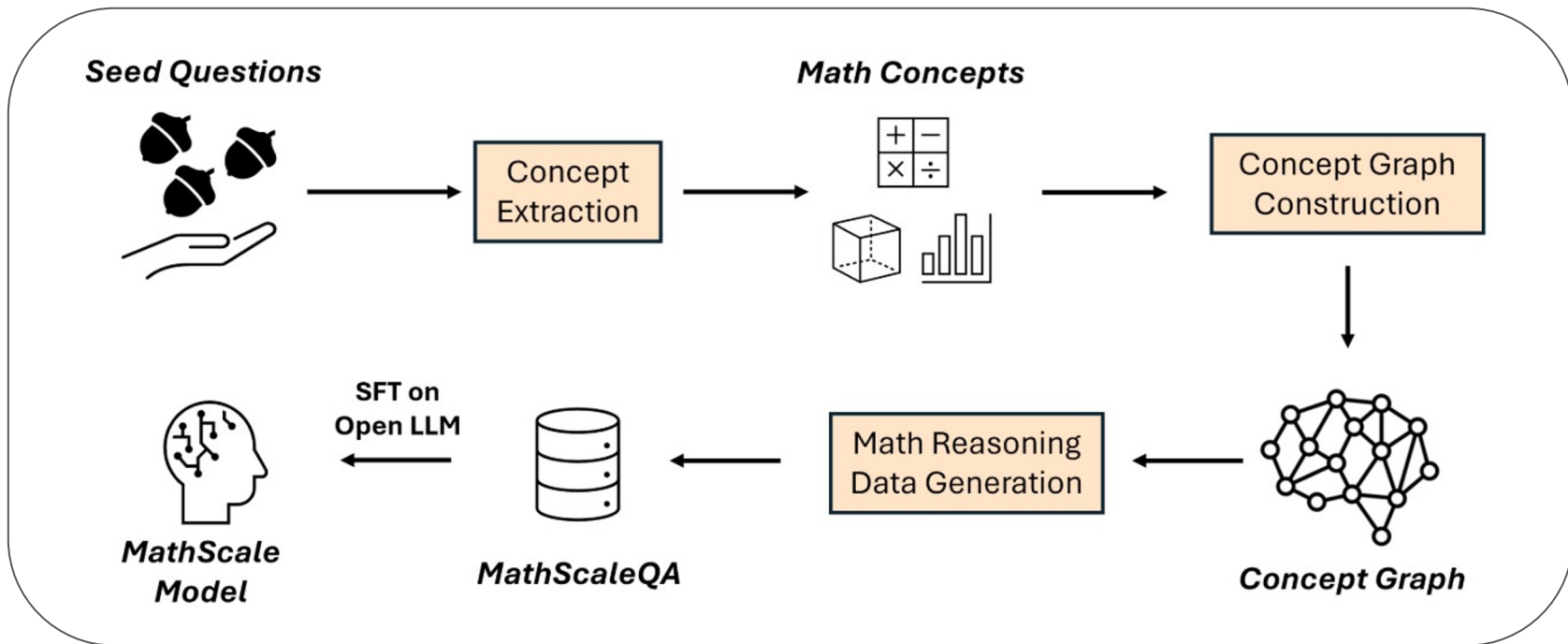
MathScale: towards Scaling Instruction Tuning for Math

Mathematical Learning Process:

(1) Concept Compression

(2) Connection Forging

[\(Tall, David, 2013\)](#)



MathScale: Concept Extraction

(1) Concept Extraction Prompt

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:

(2) Examples of Extracted Topics

"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"

(3) Examples of Extracted Knowledge Points

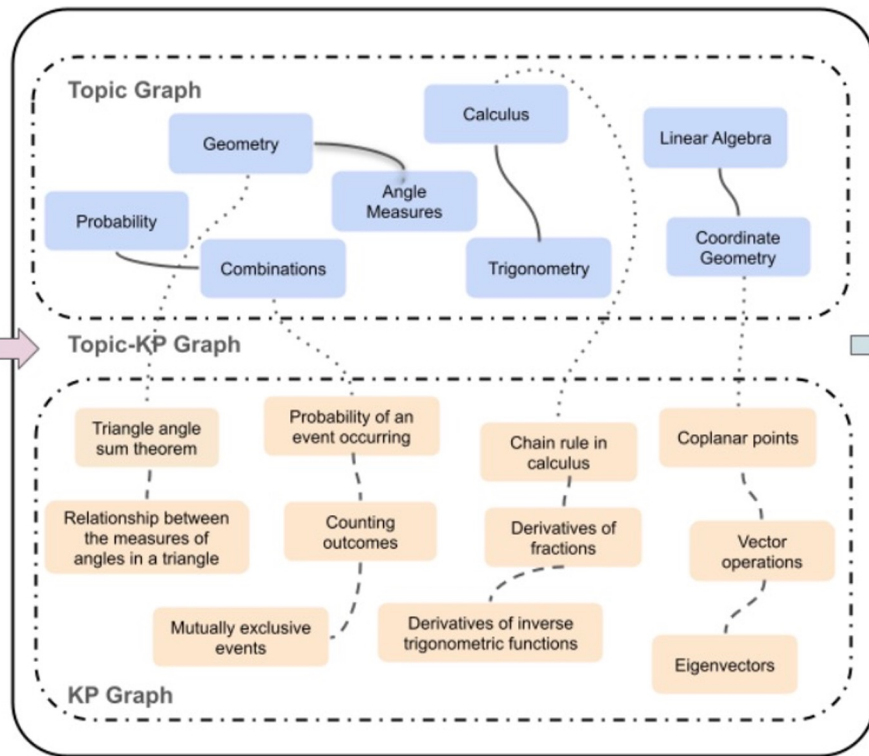
"Random selection of marbles" "Definition and properties of dot product" "Manipulation of complex numbers"
"Calculation of time required to complete a task" "How to apply the concept of a seven-day cycle" "Distinct numbers"
"Expectation of a function of a random variable" "Ability to calculate total time" "Combinations of numbers"

MathScale: Concept Graph Construction

Seed Questions

Q1: Let the first angle of a triangle be x degrees. Find the measures of the second and third angles.

Q2: If three coins are tossed at the same time, what is the probability of getting two tails and one head? Express your answer as a common fraction.
...



Generated Questions

Q1: If the degree measures of the angles of a triangle are in the ratio 3: 4: 5, what is the degree measure of the largest angle of the triangle?

Q2: A bag contains 10 red marbles and 15 blue marbles. Two marbles are drawn from the bag without replacement. What is the probability that both marbles are red?
...

MathScale: Mathematical Reasoning Data Generation

(1) Data Generation Prompt

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 ...

(2) An Example of Generated Data

-----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:

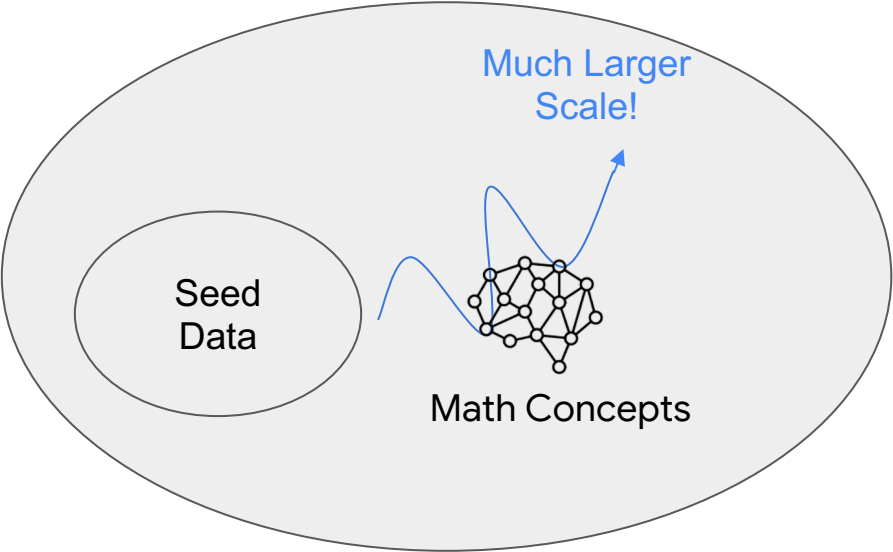
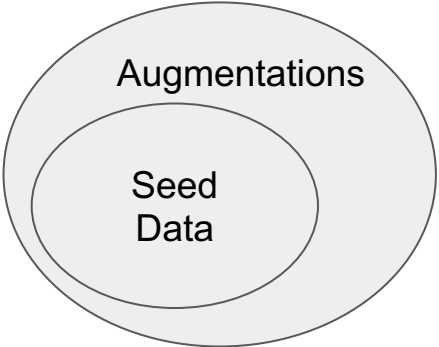
$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -4 \\ 0 & 0 & -15 \end{pmatrix}$$

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.

(3) What about validation of the generated data?

Dependency on Seed Data



MathScale: Main Results

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

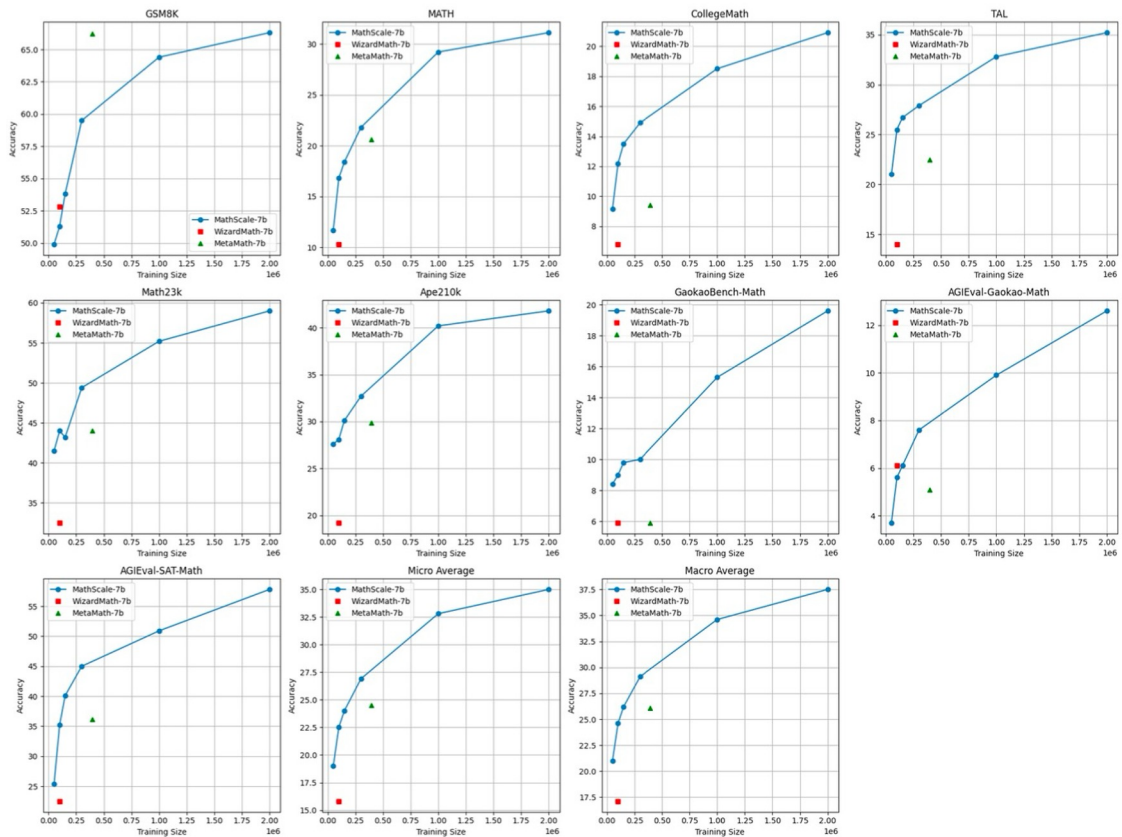
MathScale: we prompt gpt-3.5-turbo-0613 for the whole pipeline of MathScale, resulting in 2M training data points (MathScaleQA). We wrap questions in alpaca template as prompt, and treat answers as completion. We then finetune open-source LLMs on it.

Evaluation: we introduce [MWPBench \(Math Word Problem Bench\)](#) including 9 datasets across K12, college and competition math.

MathScale: More Comparisons

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

Scaling Property of MathScale



MathScale: More Ablations

(1) Ablations of MathScale Pipeline

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%

(2) Performances on a Fresh Math Dataset

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

Thanks.

Twitter: @zhengyang_42

Github: <https://github.com/microsoft/unilm/tree/master/mathscale>