

MuMath: Multi-perspective Data Augmentation for Mathematical Reasoning in Large Language Models

Weihao You^{1†}, Shuo Yin^{12‡§}, Xudong Zhao^{13§}, Zhilong Ji^{1*}, Guoqiang Zhong², Jinfeng Bai¹

¹Tomorrow Advancing Life

²College of Computer Science and Technology, Ocean University of China

³School of Economics and Management, East China Jiaotong University

youweihao@tal.com, shuoyinn@foxmail.com, superazxd@163.com,

jizhilong@tal.com, gqzhong@ouc.edu.cn, jfbai.bit@gmail.com

Abstract

Recently, the tool-use Large Language Models (LLMs) that integrate with external Python interpreters have significantly enhanced mathematical reasoning capabilities for open-source LLMs. However, these models fall short in demonstrating the calculation process, which compromises user-friendliness and understanding of problem-solving steps. Conversely, while tool-free methods offer a clear display of the problem-solving process, their accuracy leaves room for improvement. These tool-free methods typically employ a somewhat narrow range of augmentation techniques such as rephrasing and difficulty enhancement to boost performance. In response to this issue, we have amalgamated and further refined these strengths while broadening the scope of augmentation methods to construct a **multi-perspective** augmentation dataset for **mathematics**—termed **MuMath** (μ -Math) Dataset. Subsequently, we finetune LLaMA-2 on the MuMath dataset to derive the MuMath model. Our experiments indicate that our MuMath-70B model achieves new state-of-the-art performance among tool-free methods—achieving 88.3% on GSM8K and 34.5% on MATH. We release the MuMath dataset along with its corresponding models and code for public use.

1 Introduction

Large Language Models (LLMs) (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Brown et al., 2020; Raffel et al., 2023), especially proprietary LLMs like GPT-4 (OpenAI, 2023b), have been proven to be predominant across almost all the tasks in Natural Language Processing (NLP), including text classification (Jiang et al., 2023b; Min et al., 2022), code generation (Chen et al., 2021;

[†]Equal contribution.

[§]Work done while the author was interning at TAL.

^{*}Corresponding author.

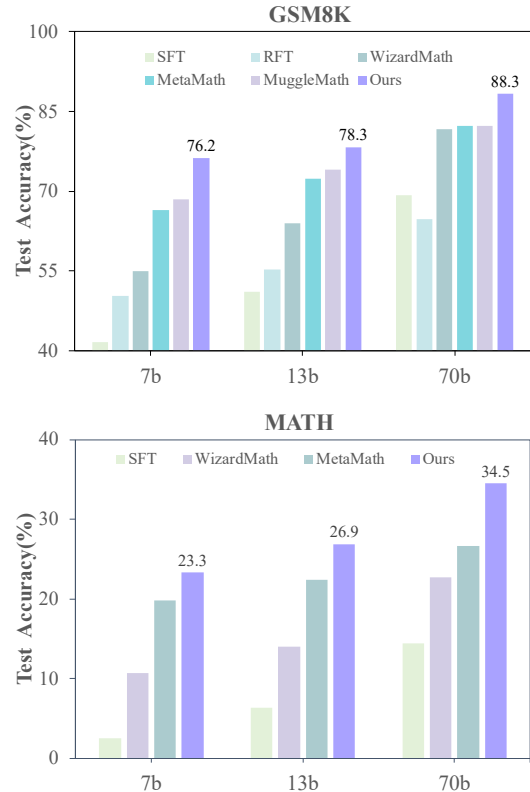


Figure 1: Comparing MuMath with baselines on LLaMA-2 base models from 7B to 70B, it’s observed that MuMath demonstrate significant enhancement over previous state-of-the-art mathematical reasoning LLMs.

Luo et al., 2023b), instruction following (Longpre et al., 2023), and mathematical reasoning (Li et al., 2023; Yu et al., 2023; Gou et al., 2023). Among these, mathematical ability is an important and typical aspect for evaluating different LLMs, and there still remains a considerable gap between open-source LLMs, e.g., LLaMA (Touvron et al., 2023), and the proprietary LLMs in the realm of mathematical problem solving (Yue et al., 2023).

Recently, a multitude of studies dedicated to enhancing the mathematical capabilities of open-source LLMs, which can be generally divided into two different research trajectories: tool-use

and tool-free. As for the tool-use LLMs, they are typically integrated with external Python interpreters, making full use of the latter’s impeccable abilities in numerical calculation and logical inference which can substantially assist LLMs in solving complex mathematical problems, e.g., PAL (Gao et al., 2023), PoT (Chen et al., 2023), MAMmoTH (Yue et al., 2023), TORA (Gou et al., 2023) and MathCoder (Wang et al., 2023).

Although the tool-use method can solve computational errors through code, it lacks a demonstration of the calculation process, making it less user-friendly in terms of understanding the problem-solving steps. On the other hand, while the tool-free method provides a good display of the problem-solving process, its accuracy still needs to be improved. Therefore, our work follows along the tool-free trajectory, focusing on improving the math reasoning ability of LLMs.

Representative tool-free methods adopt supervised finetuning (SFT) on the augmented datasets to enhance the LLMs’ mathematical reasoning capability, including RFT (Yuan et al., 2023), MetaMath (Yu et al., 2023), WizardMath (Luo et al., 2023a), and MuggleMath (Li et al., 2023), etc. RFT only augments the answer via rejection sampling to produce diverse reasoning paths with correct answers, but the generated data is similar to training dataset. MetaMath utilizes two simple augmentation methods, that one uses rephrasing to enhance the narrative diversity of the questions and answers, and the other adopts the SV (Weng et al., 2023) and FOBAR (Jiang et al., 2023a) to generate new mathematical problems and problem-solving strategies for equations. Instead of rephrasing, WizardMath and MuggleMath create new questions via rephrasing and difficulty enhancement, thus apparently improving the diversity of the dataset. However, the augmenting perspectives of these two methods are not sufficiently comprehensive, and the accuracy rate of the answers to new questions is suboptimal.

While their constructed augmented dataset enhances the capability of the model, different works adopt different methods and employ a rather limited variety of augmentation methods. So we integrate and further enhance their strengths and expand the perspective of augmentation methods to construct a **multi-perspective** augmentation dataset for **math**, called **MuMath** (μ -Math) Dataset, including four categories. (1) In Data Reformulation, besides the question rephrasing, we propose the solution reorganization to provide a compre-

hensive roadmap for the process and detailed answers. (2) In Backward Creation, We have retained the FOBAR method and introduced the Backward-Forward Transformation (BF-Trans) approach, which transforms equation-solving into arithmetic problem-solving, generating new problems and solution methods that are distinctly different from the FOBAR style. (3) We’ve further refined the existing question alteration from a fresh perspective: expression replacement. It offers a controllable and innovative way, compared to simply changing numbers or arbitrarily increasing difficulty. Also, we utilize majority sampling finetuning to boost answer accuracy and data quality. (4) Additionally, beyond data augmentation for mathematical problem solving, we propose a Nested Multi-task Construction Augmentation, where we nest plan programming or question summarizing texts into the solution, combining data of auxiliary tasks into the main task as solving the math problem.

Through the process of supervised fine-tuning on open-source language models, such as LLaMA-2, and applying it to the MuMath dataset, we have successfully developed MuMath models in a variety of sizes. This demonstrates that the dataset has the potential to significantly enhance the mathematical capabilities of open-source models.

Our contributions are as follows:

- We propose new data augmenting methods for math reasoning: Reorganization, BF-Trans, Expression Replacement and Nested Multi-task Construction.
- We construct a multi-perspective dataset for math, called MuMath Dataset, including data reformulation, backward creation, question alteration and nested multi-task.
- We conducted extensive experiments to demonstrate the effectiveness of different augmentations, as well as give some insights on mathematical reasoning for LLMs.
- By supervised fine-tuning on the open-source LLMs on the MuMath dataset, we obtain the MuMath model, which achieves new state-of-the-art performances among tool-free methods. MuMath-70B has achieved 88.3% on GSM8K (Cobbe et al., 2021) and 34.5% on MATH (Hendrycks et al., 2021a).

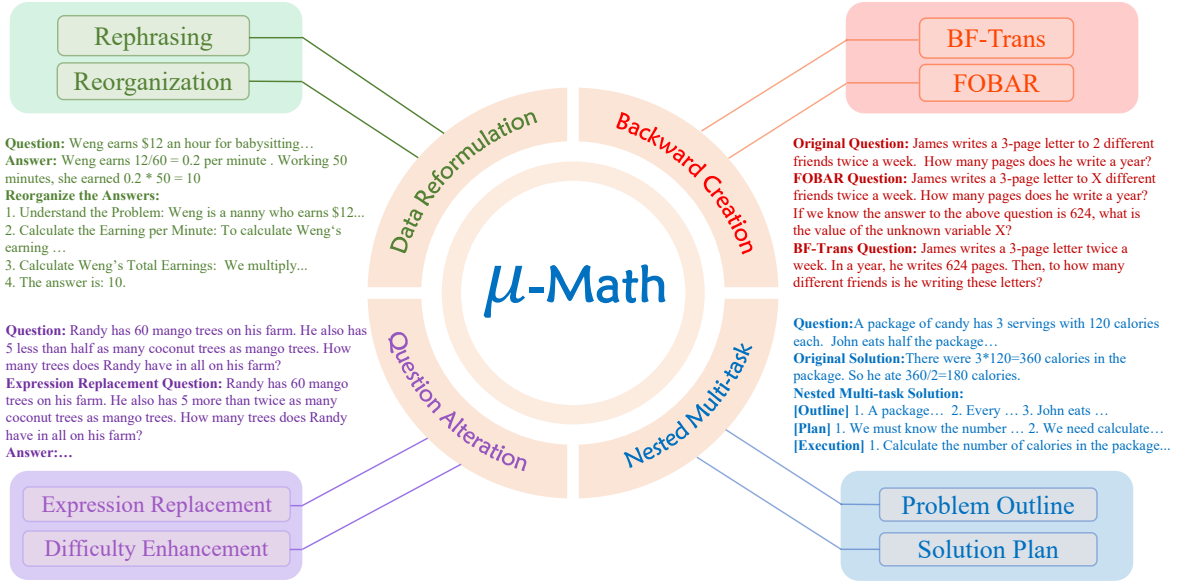


Figure 2: Overview of the augmentation methods our MuMath employs, which can be divided into four categories: (1) Data Reformulation includes solution reorganization and question rephrasing; (2) Backward Creation includes Backward-Forward Transformation (BF-Trans) and FOBAR; (3) Question Alteration includes expression replacement and difficulty enhancement; (4) Nested Multi-task construction includes data of the auxiliary tasks, i.e., Problem Outline and Solution Plan. Please zoom in the image for a better view.

2 Related Work

Mathematical Reasoning Currently, there are two main research trajectories to enhance the mathematical ability of open-source models. (1) The first trajectory focuses on LLMs purely, without tool use. Yuan et al. (2023) propose a representative tool-free methods, leveraging rejection sampling finetuning (RFT) to enhance Llama’s mathematical ability, while WizardMath (Luo et al., 2023a) chooses a reinforcement learning (RL) framework and evolves its math capability through proximal policy optimization (PPO, Schulman et al., 2017). The most recent tool-free methods are MuggleMath (Li et al., 2023) and MetaMath (Yu et al., 2023), both of which manage to augment math problem-solution data followed by finetuning the open LLMs on these newly acquired data. (2) The second trajectory underscores the integration of LLMs with tool use, with Program-aided Language model (PAL, Gao et al., 2023) and Program of Thought (PoT, Chen et al., 2023) being two pioneering works. Besides, MAMmoTH (Yue et al., 2023) employs both CoT and PoT that are combined in a coarse-grained manner, with different samples utilizing different approaches. ToRA (Gou et al., 2023) performs imitation learning with problem solutions organized by interleaving multiple steps of

program-based tool use and natural language reasoning. Similarly, MathCoder (Wang et al., 2023), contemporaneous with TORA, also curates the reasoning text in an interleaving manner, while it introduces interpolation problems to bridge the gap in difficulty between GSM8K (Cobbe et al., 2021) problems and MATH (Hendrycks et al., 2021b) problems. Our work concentrates on fully exploiting LLM’s pretrained capability to conduct mathematical reasoning, thus progressing along the tool-free research trajectory.

Data Augmentation The process of data augmentation for mathematical reasoning can be divided into two categories. The first one involves enhancing the content of both the problem and its solution without altering their original meaning. CoT approaches (Wei et al., 2022; Fu et al., 2022) present the solution in a step-by-step format, making it easier for the model to learn. Rephrasing (Yu et al., 2023; Li et al., 2023) refers to restating the problem and its solution in a different manner, thereby generating a new sequence of tokens. Rejection sampling (Yuan et al., 2023) utilizes a fully trained model to yield new reasoning paths, and the quality of these paths is intrinsically linked to the performance of the trained model. The other one modifies the computational values

Example 3.1: Reorganization

Question: Weng earns \$12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Answer: Weng earns $12 / 60 = 0.2$ per minute . Working 50 minutes, she earned $0.2 * 50 = 10$.

Reorganization Answers:

1. **Understand the Problem:** - Weng is a babysitter who earns \$12 per hour. Yesterday, she worked for 50 minutes. Our task is to calculate how much she earned.

2. **Calculate the Earning per Minute:** - To calculate Weng’s earning per minute, we divide her hourly earning by 60 (the total number of minutes in an hour): $12 / 60 = 0.2$ per minute.

3. **Calculate Weng’s Total Earnings:** - We multiply Weng’s earning per minute by the number of minutes she worked: $0.2 * 50 = 10$.

4. **Answer:** - Weng earned 10 for her 50 minutes of babysitting. **The answer is: 10.**

or logic of a problem, thus generating new problems. Then, through the method of knowledge distillation (Huang et al., 2022; Li et al., 2022; Magister et al., 2023; Ho et al., 2023; Fu et al., 2023b; Shridhar et al., 2023), it generates new solutions and transfers reasoning abilities from the teacher model (for instance, GPT4). The Evol-instruct method (Xu et al., 2023; Luo et al., 2023a) and difficulty enhancement (Li et al., 2023) incorporate modifications such as adding constraints, adjusting the context, and more to the original data. FOBAR (Jiang et al., 2023a) generates a series of questions for backward reasoning by masking numbers. It then samples a set of backward reasoning chains to predict the masked number. Our proposed method not only enriches these two types of augmentation, but also adds a multi-task augmentation category. This can be nested into the existing data to bolster the mathematical reasoning capabilities of the model.

3 Methods

The overview of our method is illustrated in Figure 2. The implementation of our proposed data augmentation methods is to request GPT4 to obtain the desired data through specific prompts.

3.1 Data reformulation

Our data reformulation can be divided into two primary categories: rephrasing and reorganization.

Rephrasing Rephrasing refers to rewriting a text while keeping the original meaning unchanged, which is also used in MetaMath (Yu et al., 2023). The prompt we use for rephrasing is shown in Prompt B.1. After requesting answers to the rephrasing questions, we can get $\mathcal{D}_{reph} = \{(Q_{reph}, S_{reph})\}$ by filtering out questions with incorrect answers.

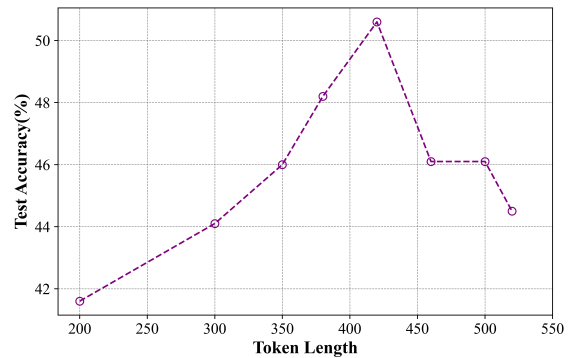


Figure 3: The relationship between token length and accuracy on GSM8K test set.

Reorganization While rephrasing augments questions without altering the original meaning, reorganization merely amplifies a solution which also holds the same meaning as the original solution. We believe that solutions which are both standardized and detailed tend to be more easily comprehended. So we have made solving steps more understandable for learning by reorganization. After the reorganization through the LLM, the solving steps will be more logically organized and clearer. Phrases such as "understand the problem", "define variables", and "calculate the number" act as explicit instructions, leading us toward the final result by "The answer is". See Example 3.1 for details. The prompt we use for reorganization is shown in Prompt B.2. We use S_{reorg} to denote the reconstructed solution, and thus the new dataset we get can be formalized as $\mathcal{D}_{reorg} = \{(Q, S_{reorg})\}$.

For the reorganization solutions, we manipulated response length by adding a minimum word count restriction in the prompt. Upon examining the generated response, it was discerned that longer token lengths corresponded to lower complexity in overall responses. However, the parsing steps become redundant when the token length becomes exces-

sively long. The result could potentially lead to models assimilating irrelevant information while overlooking correct answers. See the example in Appendix A. Consequently, this underscores the importance of optimal response length for ensuring model efficacy during reorganization augmentation. So we fine-tune LLaMA-2 7B utilizing data of varying token lengths and subsequently depict the correlation between token length and accuracy. Figure 3 shows a linear accuracy increase for token lengths between 200 and 420, but the accuracy begins to decline when the token length exceeds 420. So we have chosen to utilize a token length of approximately 420 for the reorganization data.

Combining the rephrasing and reorganization datasets, we have the reformulation dataset $\mathcal{D}_1 = \mathcal{D}_{reorg} \cup \mathcal{D}_{reph}$.

3.2 Backward-Forward Transformation

FOBAR (Jiang et al., 2023a) masks some specific value in the original forward question using “X”, convert the final answer to a new condition, and thus construct a backward question by asking to find the unknown variable X. However, this method tends to list equations concerning X and then solve them, as still a forward reasoning process. Here our purpose is to introduce backward questions with directly arithmetic solutions instead of equation solving, i.e., engage in as much reverse reasoning as possible.

To this end, we propose a new method called **Backward-Forward Transformation (BF-Trans)**. For a certain question-answer pair, we firstly utilize FOBAR to transform the original question Q into a backward one Q_b ; secondly, we rephrase the FOBAR question into a new form where the masked value is requested directly instead of employing an unknown variable X, resulting in a “secondary forward” question which we called BF-Trans question, marked as Q_{bf} . Example 3.2 shows the differences among the original question, FORAR and BF-Trans. Finally, we generate the solution S_{bf} for this BF-Trans question. Collecting all these BF-Trans augmented samples, we can have $\mathcal{D}_{bf} = \{(Q_{bf}, S_{bf})\}$. Note that the final answer of the BF-Trans solution is correct after the filtering procedure, corresponding to a certain masked number of the FOBAR question is corresponding to a certain number. See Prompt B.3 and B.4 for more details.

Combined with the FOBAR dataset \mathcal{D}_{fobar} , hence the backward reasoning part of our final train-

ing set is $\mathcal{D}_2 = \mathcal{D}_{bf} \cup \mathcal{D}_{fobar}$.

3.3 Question Alteration

Our observations have highlighted that diversity and complexity inherent within training data play an instrumental role in enhancing mathematical reasoning capabilities. So we also strive to enhance our model’s ability to generalize by generating brand new problems. We have employed a more diversified perspective in generation and significantly enhanced the quality of our data.

Difficulty Enhancement Drawing inspiration from WizardMath (Luo et al., 2023a) and MuggleMath (Li et al., 2023), we increase the problem difficulty to create new questions $Q_{complex}$. Our methods include but are not limited to adding constraints and modifying context. The prompt we use for getting more difficult questions are in Prompt B.5.

Expression Replacement We assert that changing numerals doesn’t alter the logic of the calculation, representing a singular enhancement. Conversely, arbitrarily increasing difficulty is excessively unrestricted. Thus, to broaden our perspective on question alteration, we introduce expression replacement as a novel and controllable alteration method that has a different calculation logic intrinsically. This method offers an interpolated perspective between changing numerical and increasing difficulty arbitrarily. The comparison of these three methods can be found in the Appendix A. Our approach operates in this manner: we first extract all mathematical expressions from the solution. Subsequently, an arithmetic expression is altered to form a novel equation. With the original problem statement and new equations as guides, a new question can be generated denoted as $Q_{replace}$. Example 3.3 compares the original questions and the expression replaced one. Refer to Prompt B.6 and B.7 for more details about how we get the new questions.

The newly generated questions via expression replacement may indeed be unreasonable and unsolvable. Therefore, when requesting answers, we utilize GPT-4 to act as a judge model, assessing the validity of the new questions. Only those questions deemed reasonable are requested for answers and added to the training set.

Majority Sampling Finetuning After generating new questions, we utilized GPT4 for solutions. A challenge emerges as these new questions lack

Example 3.2: FOBAR vs BF-Trans

FOBAR Question: Joe’s mother gave him \$56 to go to the store. Joe bought X notebooks and 2 books. Each notebook costs \$4 and each book costs \$7. How much money does Joe have left? If we know the answer to the above question is 14, what is the value of the unknown variable X ?

FOBAR Response: We know Joe have \$14 left after he bought X notebooks and 2 books. Each notebook costs \$4 and he bought X of them, so he spent $4X$ dollars on notebooks. Each book costs \$7 and he bought 2 of them, so he spent $2 \times 7 = 14$ dollars on books. Hence the total money he spent is $4X + 14$. We know he originally had \$56, so we have $56 - (4X + 14) = 42 - 4X$ left. Since we know the answer is 14, we have $42 - 4X = 14$. **Solving for X ,** we get $4X = 28$ and $X = 7$.

BF-Trans Question: Joe’s mother gave him \$56 to go to the store. Joe bought 2 books and some notebooks. Each notebook costs \$4 and each book costs \$7. If Joe has \$14 left after shopping, how many notebooks did he buy?

BF-Trans Response : Joe initially had \$56, and now has \$14 left after shopping, so he spent $56 - 14 = 42$ dollars. Each book costs \$7 and he bought 2 books, so he spent $7 \times 2 = 14$ dollars on books. He spent 42 dollars in total and 14 dollars on books, so he spent $42 - 14 = 28$ dollars on notebooks. Each notebook costs \$4, so he bought $28 / 4 = 7$ notebooks.

Example 3.3: Expression Replacement

Question: Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Response: Half of the number of Randy’s mango trees is $60 / 2 = 30$ trees. So Randy has $30 - 5 = 25$ coconut trees. Therefore, Randy has $60 + 25 = 85$ trees on his farm. The answer is: 85

New Question: Randy has 60 mango trees on his farm. He also has 5 more than twice as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

New Response: Twice the number of mango trees on Randy’s farm is $60 \times 2 = 120$ trees. The total number of coconut trees on Randy’s farm is 5 more than twice the number of mango trees, a total of $120 + 5 = 125$ trees.

Altogether, Randy has $125 + 60 = 185$ trees on his farm. The answer is: 185

standard reference answers, possibly introducing errors into the training data. Despite this, our experiments showed satisfactory performance from models trained with this data. We hypothesize that correct steps within incorrect final answers might assist LLMs in understanding math problems, aligning with theories proposed in (Fu et al., 2023a) and (Yu et al., 2023). To maximize answer accuracy for new questions, we implemented Majority Solution Sampling to achieve a higher-accuracy dataset for these queries. We utilize majority voting with $k = 30$ to request solutions and only select one response with the majority answer for finetuning. We name the above procedure as Majority Sampling Finetuning (MSF).

We use $S_{replace}$ and $S_{complex}$ to stand for the generated solutions to the newly introduced questions $Q_{replace}$ and $Q_{complex}$ respectively, resulting in our recreation dataset $\mathcal{D}_3 = \{(Q_{replace}, S_{replace})\} \cup \{(Q_{complex}, S_{complex})\}$.

3.4 Nested Multi-task Learning

Multitask learning (Raffel et al., 2023; Sun et al., 2019) equips a single model with the capability to handle diverse tasks, and it can also enhance the main task processing ability of the model, by introducing strongly correlated auxiliary tasks. Different from continual learning (Parisi et al., 2019)

where different tasks are separated in stage level (thus coarse-grained), multitask learning is a fine-grained procedure, and it integrates the data from different tasks into a single training batch for simultaneously learning (different tasks are distinguished in batch level). We propose a more fine-grained multi-task learning strategy called Nested Multi-Task learning (NestedMT), where we nest the data of auxiliary tasks into the data of the main task in a sample level.

Specifically, for the main task of solving mathematical problems Q , we select two auxiliary tasks: summarizing the question and listing the solving plan. Different from the stage-level and batch-level counterparts, we prepend the text of question outline O , solving plan P , or both to the solution text S , assembling into an individual final solution $S_{mt} = O \oplus P \oplus S$, where \oplus represents concatenation, for each original question. More details are shown in Example A.3 and Prompt B.8. Then we have $\mathcal{D}_4 = \{(Q, S_{mt})\}$ as the nested multi-task dataset. In nested multi-task learning, our model can learn to solve the math problems and meanwhile learn to manage various auxiliary tasks strongly related to the math problem solving task itself. All these tasks are concentrated into one single sample and thus the auxiliary tasks can contribute in a more detailed and precise manner to

improve the model’s performance on its principal task as math problem solving.

4 Experiments

4.1 Experimental Setup

Datasets We employ two widely recognized mathematical reasoning benchmarks. The first one, GSM8K (Cobbe et al., 2021), is a collection of high-quality elementary school math problems, comprising 7,473 training instances and 1,319 test instances. The second benchmark is the MATH dataset (Hendrycks et al., 2021a), which encompasses seven subjects, i.e., Prealgebra, Algebra, Number Theory, Counting and Probability, Geometry, Intermediate Algebra and Precalculus. This dataset includes math competition problems from high school level with a total of 7,500 training samples and 5,000 testing samples.

We employ a series of augmentation methods mentioned in Section 3, to create different subsets based on the original GSM8K and MATH training data. Note that there are significant differences in difficulty levels and numbers of conditions between questions of these two datasets. Therefore, after requesting new solutions and the subsequent filtering, the amounts of data we obtained from GSM8K and MATH are slightly different.

For question augmentation, we firstly employ rephrasing, alteration, FObAR and BF-Trans to get about 7k questions for each method on each original dataset. Then we make multiple requests for solutions to all these questions (15 times on GSM8K, and 30 times on MATH). We use majority voting to select samples for data augmented via alteration, which have no ground truth answer; for the other parts (rephrasing, FObAR and BF-Trans), we filter out samples with wrong answers. After that, we vary the maximum number of samples for one unique question (denoted as n), and plot the accuracy curves of the 7B models tested on GSM8K and MATH (see Appendix C). We select a point ($n = 2$) with an appropriate amount of data and relatively strong performance, then proceed to sampling, and finally obtain the subsets of the resulting MuMath dataset (about 277K). Reorganization and Nested Multi-task Construction are merely solution augmentation conducted on the original data, about 7K for each method on each dataset (totally 27K). These above subsets add up to our final MuMath dataset (304K).

Implementation Details Our study utilizes the state-of-the-art open-source LLMs for fine-tuning, comprising LLaMA-2 7B, LLaMA-2 13B, and LLaMA-2 70B (Touvron et al., 2023). All these models undergo full fine-tuning. We incorporate system prompts from (Taori et al., 2023) during the fine-tuning, and employ AdamW for optimization. We set the global batch size to 128 and used a cosine learning rate scheduler with a 0.03 warm-up period for 3 epochs. The computational hardware are NVIDIA A800 GPUs.

Model	GSM8K	MATH
<i>colsed-source LLMs</i>		
GPT-4 (OpenAI, 2023b)	92.0	42.5
GPT-3.5-Turbo (OpenAI, 2023a)	80.8	34.1
PaLM (540B)(Chowdhery et al., 2022)	56.5	8.8
PaLM-2 (540B) (Anil et al., 2023)	80.7	34.3
Minerva (540B) (Lewkowycz et al., 2022)	58.8	33.6
<i>tool-use LLMs</i>		
<i>7B</i>		
CodeLLaMa (PAL) (Rozière et al., 2023)	34.0	16.6
MAmmoTH (Yue et al., 2023)	53.6	31.5
MathCoder-L (Wang et al., 2023)	64.2	23.3
ToRA (Gou et al., 2023)	68.8	40.1
<i>13B</i>		
CodeLLaMa (PAL) (Rozière et al., 2023)	39.9	19.9
MAmmoTH (Yue et al., 2023)	62.0	34.2
MathCoder-L (Wang et al., 2023)	72.6	29.9
ToRA (Gou et al., 2023)	72.7	43.0
<i>70B</i>		
MAmmoTH (Yue et al., 2023)	76.9	41.8
MathCoder-L (Wang et al., 2023)	83.9	45.1
ToRA (Gou et al., 2023)	84.3	49.7
<i>tool-free LLMs</i>		
<i>7B</i>		
LLaMA-2 (Touvron et al., 2023)	14.6	2.5
LLaMA-2 SFT (Touvron et al., 2023)	41.6	-
LLaMA-2 RFT (Yuan et al., 2023)	50.3	-
WizardMath (Luo et al., 2023a)	54.9	10.7
MetaMath† (Yu et al., 2023)	66.3	19.7
MuggleMath (Li et al., 2023)	68.4	-
MuMath	76.2	23.3
<i>13B</i>		
LLaMA-2 (Touvron et al., 2023)	24.3	6.3
LLaMA-2 SFT (Touvron et al., 2023)	51.1	9.2
LLaMA-2 RFT (Yuan et al., 2023)	55.3	-
WizardMath (Luo et al., 2023a)	63.9	14
MetaMath (Yu et al., 2023)	72.3	22.4
MuggleMath (Li et al., 2023)	74	-
MuMath	78.3	26.9
<i>70B</i>		
LLaMA-2 (Touvron et al., 2023)	57.8	14.4
LLaMA-2 SFT (Touvron et al., 2023)	69.3	14.9
LLaMA-2 RFT (Yuan et al., 2023)	64.8	-
WizardMath (Luo et al., 2023a)	81.6	22.7
MetaMath (Yu et al., 2023)	82.3	26.6
MuggleMath (Li et al., 2023)	82.3	-
MuMath	88.3	34.5

Table 1: Comparison of testing accuracy to existing LLMs on GSM8K and MATH. † denotes the results are our own reproduction of MetaMath 7B (finetuned on MetaMathQA), which are close to the ones in the original paper.

4.2 Comparison Results

In Table 1, we contrast the performance of current colsed-source LLMs, tool-use LLMs, and tool-free LLMs on GSM8K and MATH. It’s evident that MuMath set a new standard in the 7B LLMs. Compared to the baseline LLaMA-2 SFT, MuMath shows significant accuracy increases on GSM8K and MATH by 34.6% and 18.9%, respectively. In contrast to MetaMath, MuMath improves by 9.9% and 3.6% on GSM8K and MATH respectively. In LLMs with 13B parameters, MuMath surpasses MetaMath by 6% and 4.5% on GSM8K and MATH datasets respectively. For LLMs with 70B parameters, MuMath surpasses MetaMath by 6% on the GSM8K dataset. Significantly, against MetaMath on the MATH dataset, MuMath improves impressively by a margin of 7.9%. Note that our MuMath dataset contains approximately **304K** samples, apparently less than that of MetaMathQA (**395K**). This highlights our proposed data augmentation methods’ effectiveness in enhancing mathematical reasoning capabilities.

4.3 Ablation of Different Augmentation

In this section, we conduct experiments to study the effect of augmentations in MuMath. Table 2 shows the fine-tuning results of each sub-component within our proposed augmentation methods, tested on both GSM8K and Math datasets. The data size of each subset is consistent with the original data (7K). Each dataset shows substantial improvement compared to the original data. Remarkably, the nested multi-task augmentation records a 9.4% increase under equal quantities on GSM8K. To sum up, all of our augmentation methods effectively boost the mathematical reasoning abilities of open-source LLMs.

Method	GSM8K		MATH	
	Datasize	Acc	Datasize	Acc
SFT	7K	41.6	7K	4.4
Reorganization	7K	50.6	7K	6.0
Rephrasing	7K	46.2	7K	5.9
Reorganization + Rephrasing	7K+7K	52.1	7K+7K	7.3
FOBAR	7K	40.6	7K	4.9
BF-trans	7K	42.8	7K	5.8
FOBAR + BF-Trans	7K+7K	46.2	7K+7K	7.4
Expression Replacement (ER)	7K	47.7	7K	6.4
Complexity Enhancement (CE)	7K	45.1	7K	4.6
ER + CE	7K+7K	48.5	7K+7K	7.0
Nested Multi-task	7K	51.0	7K	6.8
Separate Multi-task	7K+7K	42.5	7K+7K	6.6

Table 2: Different data augmentation strategies on GSM8K and MATH performances.

GSM8K					MATH				
\mathcal{D}_1	\mathcal{D}_2	\mathcal{D}_3	\mathcal{D}_4	Acc.	\mathcal{D}_1	\mathcal{D}_2	\mathcal{D}_3	\mathcal{D}_4	Acc.
21K	40K	74K	7K		21K	40K	94K	7K	
✓	✗	✗	✗	59.6	✓	✗	✗	✗	10.5
✗	✓	✗	✗	53.3	✗	✓	✗	✗	10.7
✗	✗	✓	✗	57.7	✗	✗	✓	✗	17.9
✗	✗	✗	✓	51.0	✗	✗	✗	✓	6.8
✓	✓	✗	✗	64.0	✓	✓	✗	✗	14.5
✓	✗	✓	✗	64.5	✓	✗	✓	✗	19.1
✓	✗	✗	✓	60.8	✓	✗	✗	✓	10.8
✗	✓	✓	✗	62.2	✗	✓	✓	✗	20.2
✗	✓	✗	✓	55.6	✗	✓	✗	✓	12.6
✗	✗	✓	✓	60.1	✗	✗	✓	✓	18.6
✓	✓	✓	✗	67.9	✓	✓	✓	✗	21.1
✓	✓	✗	✓	65.1	✓	✓	✗	✓	14.8
✓	✗	✓	✓	64.0	✓	✗	✓	✓	20.1
✗	✓	✓	✓	63.2	✗	✓	✓	✓	20.6
✓	✓	✓	✓	69.2	✓	✓	✓	✓	21.6
MetaMath				64.4					17.7
MuggleMath				68.4					-

Table 3: Effect of different data subsets on the accuracy of GSM8K and MATH. $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ and \mathcal{D}_4 are data reformulation, backward creation, question alteration, and nested multi-task learning. We also compare our MuMath model with two baselines, all of which are trained on datasets augmented from only one source.

Moreover, from the results obtained by the stacked data, we discovered that the sub-methods within each of the four data augmentation methods are complementary to each other.

Table 3 enumerates the data volumes of four augmentation datasets, and it mainly presents the test accuracy of various augmentation combinations. As observed, the models trained on any kind of augmentations outperform the SFT method significantly. In the GSM8K, employing a single data augmentation method enables data reformulation to attain an accuracy rate of 59.6%. In the MATH, using only question alteration data yields a 17.9% accuracy rate. Surprisingly, when combining multiple data augmentation methods in any manner, each additional data increment contributes to further enhancement. This phenomenon persists even at high accuracy levels. This highlights the versatility and effectiveness of each augmentation method.

4.4 MSF vs. SFT

We extract 7K new created questions from MATH to validate our proposed Majority Sampling Fine-

tuning (MSF). Specifically, for each question we randomly select n solutions with the majority answer to construct MSF dataset (for those questions with less than n majority solutions, we compromise to use all the $< n$ solutions), and directly request n solutions with different answers to construct SFT dataset. Figure 4 illustrates that as the amount of training data increases (with n varying from 1 to 8), models trained using MSF and SFT both see a progressive improvement in their performance. However, the latter saturates earlier than the former, and across all data sizes, the MSF models consistently outperform the SFT ones.

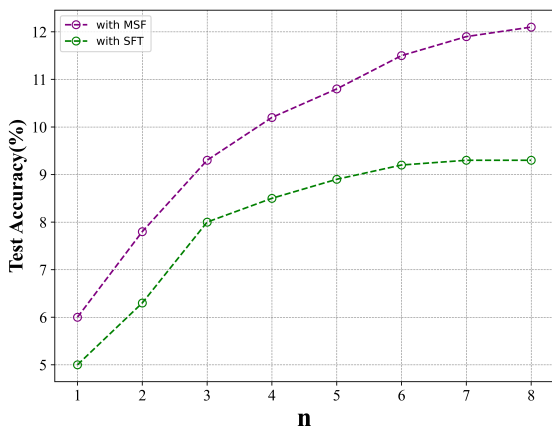


Figure 4: Comparison of performance between models trained with MSF and with SFT on MATH dataset.

4.5 Out-of-Domain Math Reasoning

We have evaluated our MuMath 7B and 70B models on out-of-domain datasets, including SVAMP, MAWPS and ASDiv. The results are shown in Table 4. On 2 out of 3 above datasets, the performances of our MuMath can even surpass the state-of-the-art tool-use open LLM, ToRA.

We conduct another ablation study to test the out-of-domain math reasoning capability of MuMath. We firstly split our MuMath dataset into 2 subsets, GSM8K augmented subset (142k) and MATH augmented part (162k). The in-domain test and out-of-domain test are shown in Table 5. Apparently the out-of-domain reasoning results are bad, consistent with the observation in MuggleMath. According to the results in Table 5 and those in Table 1, we can conclude that GSM8K augmented data do not help much in improving the accuracy on MATH and vice versa, which matches the ablation results in MetaMath.

Models	SVAMP	MAWPS	ASDiv
<i>7B</i>			
WizardMath-7B	57.3	73.3	59.1
ToRA-7B	70.4	91.3	78.7
MuMath-7B	76.8	87.3	93.6
<i>70B</i>			
WizardMath-70B	80.0	86.2	76.2
ToRA-70B	82.7	93.8	86.8
MuMath-70B	87.6	92.0	96.6

Table 4: The out-of-domain reasoning capability comparison between MuMath and the other methods.

	GSM8K (test)	MATH (test)
GSM8K (train)	69.2	6.7
MATH (train)	42.4	21.6

Table 5: The in-domain and out-of-domain reasoning capability of MuMath 7B on GSM8K and MATH.

5 Conclusion

In this work, we propose four novel methods to broaden the scope of augmentation for mathematical reasoning data: solution reorganization, BF-Trans, expression replacement and nested multi-task construction. Through a variety of augmenting strategies, we create a multi-perspective mathematical problem-solving dataset based on GSM8K and MATH, called MuMath. After finetuning LLaMA-2 on the novel dataset, we get a series of models (7B, 13B and 70B) equipped with excellent math capability, which are also termed MuMath. Extensive empirical results demonstrate the effectiveness of our proposed augmentation methods. Compared to the open-source methods, our MuMath achieves the best performance in tool-free LLMs across all model scales, and even surpasses some tool-use counterparts. We will explore other augmentation methods for further improving mathematical reasoning performance of tool-free LLMs, as well as more auxiliary tasks for nested multi-task learning.

6 Acknowledgments

This work was supported by National Key R&D Program of China under Grant No. 2020AAA0104500, HY Project under Grant No. LZ2022033004, the Natural Science Foundation of Shandong Province under Grants No. ZR2020MF131 and No. ZR2021ZD19, the Science and Technology Program of Qingdao under Grant No. 21-1-4-ny-19-nsh, and Project of Associative Training of Ocean University of China under Grant No. 202265007.

References

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#).
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023a. [Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance](#).
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023b. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Pal: Program-aided language models](#).
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujia Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. [Tora: A tool-integrated reasoning agent for mathematical problem solving](#).
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021a. Measuring mathematical problem solving with the MATH dataset.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#).
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. [Large language models can self-improve](#).

- Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James T. Kwok. 2023a. [Forward-backward reasoning in large language models for mathematical verification.](#)
- Weisen Jiang, Yu Zhang, and James Kwok. 2023b. [Effective structured prompting by meta-learning and representative verbalizer.](#) In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15186–15199. PMLR.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.
- Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2023. [Query and response augmentation cannot help out-of-domain math reasoning generalization.](#)
- Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhui Chen, and Xifeng Yan. 2022. [Explanations from large language models make small reasoners better.](#)
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach.](#)
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. [Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct.](#)
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. [Wizardcoder: Empowering code large language models with evol-instruct.](#) *arXiv preprint arXiv:2306.08568*.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason.](#)
- Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hananeh Hajishirzi. 2022. [MetaICL: Learning to learn in context.](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.
- OpenAI. 2023a. [Chatgpt: Optimizing language models for dialogue.](#) <https://openai.com/blog/chatgpt>.
- OpenAI. 2023b. [Gpt-4 technical report.](#)
- German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. [Continual lifelong learning with neural networks: A review.](#) *Neural Networks*, 113:54–71.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, , and I. Sutskever. 2019. Language models are unsupervised multitask learners. *Technical Report*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer.](#)
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. [Code llama: Open foundation models for code.](#)
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms.](#)
- Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. [Distilling reasoning capabilities into smaller language models.](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. [Ernie 2.0: A continual pre-training framework for language understanding.](#)
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford alpaca: An instruction-following llama model.](#) https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

- Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. [Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. [Large language models are better reasoners with self-verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. [Wizardlm: Empowering large language models to follow complex instructions](#).
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. [Meta-math: Bootstrap your own mathematical questions for large language models](#).
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. [Scaling relationship on learning mathematical reasoning with large language models](#).
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mammoth: Building math generalist models through hybrid instruction tuning](#).

A Examples

Example A.1: Knowledge

Question: Weng earns \$12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Answer:(200 token length)

1. Understand the Problem: - Weng is a babysitter who earns \$12 per hour. - Yesterday, she worked for 50 minutes. - Our task is to calculate how much she earned.
2. Calculate the Earning per Minute: - To calculate Weng's earning per minute, we divide her hourly earning by 60 (the total number of minutes in an hour): $\$12 / 60 = \0.2 per minute.
3. Calculate Weng's Total Earnings: - We multiply Weng's earning per minute by the number of minutes she worked: $\$0.2 \times 50 = \10 .
4. Answer: - Weng earned \$10 for her 50 minutes of babysitting.

Answer:(520 token length)

1. Analyze and Breakdown the Problem: - In this problem, Weng is earning a rate of \$12 per hour for babysitting. - However, she hasn't worked a full hour; she's only worked for 50 minutes. - Therefore, the challenge here is to figure out just how much she's earned for that 50 minutes, given that her hourly rate is \$12.
2. Convert Hours to Minutes: - Since the rate Weng is given is in hours but the time she worked is in minutes, we need to convert her hourly rate to a per minute rate to simplify the calculations. - There are 60 minutes in an hour, so we'll divide her hourly earnings by 60 to find out how much she earns per minute. - The calculation is as follows: $12/60 = \$0.2$ per minute.
3. Connect to real-life situations: - Picture Weng jotting down her babysitting time and then calculating this on her calculator, or maybe she's keeping track in an app that monitors her babysitting earnings. - Knowing that she earns \$0.2 per minute, she would simply need to multiply this rate by the number of minutes she worked.
4. Calculate Earnings: - Now that we have Weng's per minute earning rate, we can multiply this by the number of minutes she actually worked to find out how much she earned. - Here's how that calculation looks: she worked for 50 minutes, so her earnings are computed as $0.2 \times 50 = \$10$.
5. Insert human element: - Now, Weng can feel confident that she's accurately calculated her earnings. She can spend or save this money as she chooses. Perhaps she's saving her earnings for a special purchase or maybe she counts on this income for regular expenses. Regardless, her diligence in accurately calculating her earnings assures her of what she can expect in her paycheck.
6. Answer and Conclusion: - After faithfully jotting down her babysitting time, carefully calculating her earnings, and then multiplying by the number of minutes she spent babysitting, Weng has found that she has earned \$10 for her 50 minutes of babysitting. The answer is: 10.

Example A.2: The comparison of different alteration methods

Question: Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Response: Half of the number of Randy's mango trees is $60 / 2 = 30$ trees. So Randy has $30 - 5 = 25$ coconut trees. Therefore, Randy has $60 + 25 = 85$ trees on his farm. The answer is: 85

Change numbers: Randy has 60 mango trees on his farm. He also has 10 more than twice as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Response: Half of the number of Randy's mango trees is $60 / 2 = 30$ trees. So Randy has $30 - 10 = 20$ coconut trees. Therefore, Randy has $60 + 20 = 80$ trees on his farm. The answer is: 80

Expression replacement: Randy has 60 mango trees on his farm. He also has 5 more than twice as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Response: Twice the number of mango trees on Randy's farm is $60 * 2 = 120$ trees. The total number

of coconut trees on Randy's farm is 5 more than twice the number of mango trees, a total of $120 + 5 = 125$ trees. Altogether, Randy has $125 + 60 = 185$ trees on his farm. The answer is: 185

Increase complexity: Randy has 60 mango trees on his farm. If he plants 10 more mango trees and 8 more coconut trees, the ratio of the number of mango trees to coconut trees would be 21. How many trees does Randy have in all on his farm if he decides not to plant any additional trees?"

Response: If Randy plants 10 more mango trees and 8 more coconut trees, then the ratio of the number of mango trees to coconut trees would be 21. This gives: $(m + 10) / (c + 8) = 21/1$ Simplify the Equation: - Distribute the left-hand side of the equation: $m/c + (10/c) = 21$ - Substituting $m=60$, we get $60/c + (10/c) = 21$ Solve for c: - Multiply both sides by c to eliminate the denominator: $60 + 10 = 21c$ - Combine like terms $70 = 21c$ - Solve for c: $c = 70/21$ $c = 3.33$ (rounded off) Solve for m: - Substituting the value of c which is 3.33 and $m=60$ into the equation $(m + 10) / (c + 8) = 21/1$ to see if the ratio of the number of mango trees to coconut trees would be 21: $(60+10) / (3.33+8) = 6.3636$ but 21/1 is not equal to 6.3636 -We conclude that planting 10 more mango trees and 8 more coconut trees are impossible, and Randy has 60 mango trees and no more. Answer: - Therefore, Randy has 60 trees altogether on his farm if he decides not to plant any additional trees.The answer is: 60

Example A.3: Nested Multi-task Construction

Question:

A package of candy has 3 servings with 120 calories each. John eats half the package. How many calories did he eat?

Original Solution:

There were $3 * 120 = 360$ calories in the package. So he ate $360 / 2 = 180$ calories.

The answer is: 180.

Nested Multi-task Solution:

[Outline]

1. A package of candy has 3 servings.
2. Every serving contains 120 calories.
3. John eats half the package.
4. Our task is to calculate how many calories John did eat.

[Plan]

1. Calculate the total number of calories in the package.
2. Calculate the number of calories John did eat.

[Execution]

1. One serving of the candy contains 120 calories, and there are 3 servings, so the total number of calories in the package is $3 * 120 = 360$ calories.
2. John eats half of the package, which contains $360 / 2 = 180$ calories.

The answer is: 180.

B The Prompts We Use

Prompt B.1: Rephrasing

Follow the given examples:

Question: Angelo and Melanie want to plan how many hours over the next week they should study together for their test next week. They have 2 chapters of their textbook to study and 4 worksheets to memorize. They figure out that they should dedicate 3 hours to each chapter of their textbook and 1.5 hours for each worksheet. If they plan to study no more than 4 hours each day, how many days should they plan to study total over the next week if they take a 10-minute break every hour, include 3 10-minute snack breaks each day, and 30 minutes for lunch each day?

Rephrase the above question: Angelo and Melanie need to study 2 chapters in their textbook and 4 worksheets for their upcoming test. They have planned to dedicate 3 hours for each chapter and 1.5 hours for each worksheet. They can study for a maximum of 4 hours each day, taking into account 10-minute breaks every hour, 3 10-minute snack breaks per day, and 30 minutes for lunch. How many days do they need to study in total over the next week to complete their study plan?

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

Rephrase the above question: If Leah had 32 chocolates and her sister had 42, and they both consumed 35 chocolates, what is the total number of chocolates that they have left?

Question: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

Rephrase the above question: If there were initially nine computers in the server room and five more computers were added each day from Monday to Thursday, what is the current total number of computers in the server room?

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

Rephrase the above question: If Jason initially had 20 lollipops and now has 12 after giving some to Denny, how many lollipops did he give to Denny?

Question: Sam bought a dozen boxes, each with 30 highlighter pens inside, for \$10 each box. He rearranged five of these boxes into packages of six highlighters each and sold them for \$3 per package. He sold the rest of the highlighters separately at the rate of three pens for \$2. How much profit did he make in total, in dollars?

Rephrase the above question: Sam purchased 12 boxes, each containing 30 highlighter pens, at \$10 per box. He repackaged five of these boxes into sets of six highlighters and sold them for \$3 per set. He sold the remaining highlighters individually at a rate of three pens for \$2. What is the total profit he made in dollars?

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

Rephrase the above question: If there were initially 15 trees in the grove and the grove workers are planning to plant more trees today, resulting in a total of 21 trees, how many trees did the workers plant today?

Question: {}

Rephrase the above question:

Prompt B.2: Reorganization

You are a mathematics expert. Based on the provided questions and answer process, you reorganize the Solved Process to add the token length. Let's reorganize the Solved Process. During this process, You must expand the problem-solving process to approximately 420 tokens. The methods can include: 1. increasing the complexity of the problem-solving process; 2. adding extra concepts; 3. inserting a background story; 4. increasing the steps of solving the problem. The end of response needs to be: The answer is: [answer]. Here are two examples how to do it,

==== Example 1 ====

[Problem]:

Ralph is going to practice playing tennis with a tennis ball machine that shoots out tennis balls for Ralph to hit. He loads up the machine with 175 tennis balls to start with. Out of the first 100 balls, he manages to hit $\frac{2}{5}$ of them. Of the next 75 tennis balls, he manages to hit $\frac{1}{3}$ of them. Out of all the tennis balls, how many did Ralph not hit?

[Answers]:

Out of the first 100 balls, Ralph was able to hit $\frac{2}{5}$ of them and not able to hit $\frac{3}{5}$ of them, $\frac{3}{5} \times 100 = 60$ tennis balls Ralph didn't hit. Out of the next 75 balls, Ralph was able to hit $\frac{1}{3}$ of them and not able to hit $\frac{2}{3}$ of them, $\frac{2}{3} \times 75 = 50$ tennis balls that Ralph didn't hit. Combined, Ralph was not able to hit $60 + 50 = 110$ tennis balls Ralph didn't hit. The answer is: 110

[Result]:

1. Understand the Problem:

- Here, Ralph is practicing with a tennis ball machine with 175 tennis balls.
- Ralph is hitting the balls in two batches, 100 and 75.
- From each batch, we know the fraction of balls Ralph was able to hit.
- Our task is to calculate how many tennis balls Ralph did not hit.

2. Identify the Fractions that Ralph doesn't hit:

- From the first 100 balls, Ralph manages to hit $\frac{2}{5}$ of them, meaning he was not able to hit $1 - \frac{2}{5}$ or $\frac{3}{5}$ of these balls.
- Similarly, from the next 75 balls, Ralph manages to hit $\frac{1}{3}$ of these, leaving $1 - \frac{1}{3}$ or $\frac{2}{3}$ of these balls that he doesn't hit.

3. Calculate the number of Missed Balls in the First Batch:

- We multiply the total number of balls in the first batch by the fraction that Ralph doesn't hit:
 $\frac{3}{5} \times 100 = 60$ tennis balls Ralph didn't hit from the first batch.

4. Calculate the number of Missed Balls in the Second Batch:

- We multiply the total number of balls in the second batch by the fraction that Ralph doesn't hit:
 $\frac{2}{3} \times 75 = 50$ tennis balls Ralph didn't hit from the second batch.

5. Calculate Total Missed Balls:

- Now that we have calculated the number of missed balls from both the first and second batch,
- Adding these two numbers together gives us the total missed balls:
 $60 \text{ missed balls} + 50 \text{ missed balls} = 110 \text{ missed balls}$

6. Answer:

- So, out of all the tennis balls, Ralph didn't hit 110 tennis balls during his practice session.

The answer is: 110

==== Example 2 ====

[Problem]:

Mary is making ice cubes with fruit frozen in them for a cocktail party. She makes some strawberry cubes, and a number of blueberry cubes equal to 5 times the number of strawberry cubes minus 4. If she makes 116 ice cubes total, how many blueberry cubes does she make?

[Answers]: Let s be the number of strawberry cubes and b be the number of blueberry cubes. We know that $b = 5s - 4$ and $s + b = 116$. Substituting the first equation into the second, we get $s + 5s - 4 = 116$. Combining like terms, we get $6s - 4 = 116$ Adding 4 to both sides, we get $6s = 120$

Dividing both sides by 6, we get $s = 20$ Substituting this value back into the equation $b = 5s - 4$, we get $b = 96$ The answer is: 96

[Result]:

1. Define the Variables:

- Let s represent the number of strawberry cubes.
- Let b represent the number of blueberry cubes.

2. Translate the Problem into Equations:

a. From the problem, we know that the number of blueberry cubes is 5 times the strawberry cubes minus 4. This gives:

$$b = 5s - 4$$

b. We also know the total number of cubes is 116. This gives:

$$s + b = 116$$

3. Substitute and Solve:

- Substitute the expression from the first equation into the second equation to eliminate b . - This gives:

$$s + 5s - 4 = 116$$

- Combine like terms:

$$6s - 4 = 116$$

- Solve for s :

Add 4 to both sides:

$$6s = 120$$

Divide both sides by 6:

$$s = 20$$

4. Find the Number of Blueberry Cubes:

- Using the value of s we found (20), plug it into the first equation to find b :

$$b = 5(20) - 4$$

- This gives:

$$b = 96$$

5. Answer: Mary makes 96 blueberry cubes for her cocktail party.

The answer is: 96

How about this question?

[Problem]: { }

[Answers]: { }

You must expand the problem-solving process to approximately 700 tokens. The end of response needs to be: The answer is: [answer].

[Result]:

Prompt B.3: Prompt for BF-Trans GSM8K Questions

You are an experienced mathematics teacher in a grade school, and you are good at rephrase math problems.

Now you are given a math problem (marked as [Problem]) with one and only one X as the unknown variable. Your task is to rewrite or rephrase the original problem into an equivalent problem. The equivalent problem you rephrased should not contain any X s. Instead, you should ask for the correlated unknown value using a questioning tone in the last sentence of your rephrased problem. You can use more words to keep your rephrased problem expressed clearly and thoroughly, and also can add more concepts to avoid ambiguity. Here are some examples:

==== Example 1 ====

[Problem]:

Ralph is going to practice playing tennis with a tennis ball machine that shoots out tennis balls for Ralph to hit. He loads up the machine with 175 tennis balls to start with. Out of the first 100 balls, he manages to hit X of them. Of the next 75 tennis balls, he manages to hit $1/3$ of them. Out of all the tennis balls, how many did Ralph not hit? If we know the answer to the above question is 110, what is the value of the unknown variable X ?

[Rephrase]:

Ralph is going to practice playing tennis with a tennis ball machine that shoots out tennis balls for Ralph to hit. He loads up the machine with 175 tennis balls to start with, which are divided into 2 groups. In the first group there are 100 balls and the second group contains 75 ones. Of the second group of balls, Ralph manages to hit $1/3$. And out of all the tennis balls, Ralph did not hit 110. Then out of the first 100 balls, what is the proportion of the balls Ralph hit?

==== Example 2 ====

[Problem]:

In one day, 200 people visit The Metropolitan Museum of Art in New York City. Half of the visitors are residents of New York City. Of the NYC residents, $X\%$ are college students. If the cost of a college student ticket is \$4, how much money does the museum get from college students that are residents of NYC?

If we know the answer to the above question is 120, what is the value of the unknown variable X ?

[Rephrase]:

In one day, 200 people visit The Metropolitan Museum of Art in New York City. Half of the visitors are residents of New York City. If the cost of a college student ticket is \$4, and the museum gets \$120 from college students that are residents of NYC. Then of the NYC residents, what percentage is the college students?

==== Example 3 ====

[Problem]:

X years from now, John will be 3 times as old as he was 11 years ago. How old is he now? If we know the answer to the above question is 21, what is the value of the unknown variable X ?

[Rephrase]:

If we know John is 21 years old, then how many years from now will John be 3 times as old as he was 11 years ago?

==== Example 4 ====

[Problem]:

Taipei 101 in Taiwan is X feet tall with 101 floors. Suppose the first to 100th floors have height each equal to 16.5 feet, how high is the 101st floor? If we know the answer to the above question is 23, what is the value of the unknown variable X ?

[Rephrase]:

Taipei 101 in Taiwan has 101 floors. Suppose the first to 100th floors have height each equal to 16.5 feet, and the 101st floor is 23 feet. How high is the whole building?

==== Example 5 ====

[Problem]:

A fox can run at the maximum speed of X kilometers per hour. Considering the fox would run at a constant speed, what distance would he make during 120 minutes? If we know the answer to the above question is 100, what is the value of the unknown variable X ?

[Rephrase]:

Considering a fox would run at a constant speed, and he will make 100 kilometers during 120 minutes. How many kilometers per hour the fox can run?

==== Example 6 ====

[Problem]:

Ruiz receives a monthly salary of \$500. If he received a $X\%$ raise, how much will be Ruiz's new salary? If we know the answer to the above question is 530, what is the value of the unknown variable X ?

[Rephrase]:

Ruiz receives a monthly salary of \$500. If his new salary will be \$530 monthly, what percentage is the raise?

==== Example 7 ====

[Problem]:

Tom decided to send his wife X dozen roses every day for the week. How many total roses did he send? If we know the answer to the above question is 168, what is the value of the unknown variable X ?

[Rephrase]:

Tom sent his wife 168 roses totally for the week. How many dozen roses did he sent every day for the week?

==== Example 8 ====

[Problem]:

Facebook decided to award a productivity bonus to all its female employees who are mothers. This productivity bonus will total 25% of Facebook's annual earnings, which was X for the year 2020. It is known that Facebook employs 3300 employees; one-third are men, and of the women, 1200 are not mothers. How much was the bonus that each female mother employee received, assuming each one received an equal amount? If we know the answer to the above question is 1250, what is the value of the unknown variable X ?

[Rephrase]:

Facebook decided to award a productivity bonus to all its female employees who are mothers. This productivity bonus will total 25% of Facebook's annual earnings. It is known that Facebook employs 3300 employees; one-third are men, and of the women, 1200 are not mothers. Assuming each one received an equal amount, the bonus that each female mother employee received was \$1250. Then how much was the Facebook's annual earnings for the year?

==== Example 9 ====

[Problem]: { }

[Rephrase]:

Prompt B.4: Request the solutions to BF-Trans

You are an experienced mathematician. Now you are given a grade school math problem (marked as [Problem]). The task you should accomplish is to solve this problem.

You should solve the problem step by step, as thoroughly and clearly as you can by using any number of words or solution steps. Remember in the end of your solution you should output a "The answer is:" and then output the result answer number you get. Here are some examples:

==== Example 1 ====

[Problem]:

Ralph is going to practice playing tennis with a tennis ball machine that shoots out tennis balls for Ralph to hit. He loads up the machine with 175 tennis balls to start with, which are divided into 2 groups. In the first group there are 100 balls and the second group contains 75 ones. Of the second group of balls, Ralph manages to hit $1/3$. And out of all the tennis balls, Ralph did not hit 110. Then out of the first 100 balls, what is the proportion of the balls Ralph hit?

[Solution]:

Out of all the 175 tennis balls, Ralph did not hit 110, so he hit $175 - 110 = 65$ of them. Of the second group of balls containing 75 balls, Ralph manages to hit $1/3$, so in this group he was able to hit $75 \times 1/3 = 25$. Of the first 100 balls, Ralph manages hit $65 - 25 = 40$ tennis balls, so in this group the proportion of the balls Ralph hit is $40 / 100 = 2/5$. The answer is: $2/5$

==== Example 2 ====

[Problem]:

In one day, 200 people visit The Metropolitan Museum of Art in New York City. Half of the visitors are residents of New York City. If the cost of a college student ticket is \$4, and the museum gets \$120 from college students that are residents of NYC. Then of the NYC residents, what percentage is the college students?

[Solution]:

200 people visit The Metropolitan Museum of Art in New York City, half of the visitors are residents of New York City, so the number of visitors that are residents of NYC is $200 / 2 = 100$. The cost of a college student ticket is \$4, and the museum gets \$120 from college students that are residents of NYC, so there are $120 / 4 = 30$ college students that are residents of NYC.

We know the number of visitors that are residents of NYC is 100, and there are 30 college students that are residents of NYC, so of the NYC residents, there is $30 / 100 = 30\%$ college students, meaning the percentage is 30

The answer is: 30

==== Example 3 ====

[Problem]:

Taipei 101 in Taiwan has 101 floors. Suppose the first to 100th floors have height each equal to 16.5 feet, and the 101st floor is 23 feet. How high is the whole building?

[Solution]:

The first to 100th floors have height each equal to 16.5 feet, so the total height of the first to 100th floors is $16.5 \times 100 = 1650$ feet. We know the total height of the first to 100th floors is 1650 feet, and the 101st floor is 23 feet, so the whole building is $1650 + 23 = 1673$ feet. The answer is: 1673

==== Example 4 ====

[Problem]:

Considering a fox would run at a constant speed, and he will make 100 kilometers during 120 minutes. How many kilometers per hour the fox can run?

[Solution]:

The fox will make 100 kilometers during 120 minutes, and 120 minutes are $120 / 60 = 2$ hours, so he can run $100 / 2 = 50$ kilometers per hour. The answer is: 50

==== Example 5 ====

[Problem]:

Facebook decided to award a productivity bonus to all its female employees who are mothers. This productivity bonus was total 25% of Facebook's annual earnings. It is known that Facebook employs 3300 employees; one-third are men, and of the women, 1200 are not mothers. Assuming each one received an equal amount, the bonus that each female mother employee received was \$1250. Then how much was the Facebook's annual earnings for the year?

[Solution]:

It is known that Facebook employs 3300 employees and $1/3$ are men, so $1 - 1/3 = 2/3$ are women and the number of women is $3300 \times 2/3 = 2200$

Of the women, 1200 are not mothers, so there are $2200 - 1200 = 1000$ mothers. Assuming each one received an equal amount, the productivity bonus that each female mother employee received was \$1250, and we know Of the women, there are 1000 mothers, so the total productivity bonus of the mother employees received was $\$1250 \times 1000 = \$1,250,000$

We know the total productivity bonus of the mother employees received was \$1,250,000, and it's 25% of Facebook's annual earnings for the year, so Facebook's annual earnings for the year is $\$1,250,000 / 25\% = \$1,250,000 / (1/4) = \$1,250,000 \times 4 = \$5,000,000$ The answer is: 5,000,000

==== Example 6 =====

[Problem]: { }

[Solution]:

Prompt B.5: Difficulty Enhancement

I want you to act as a math teacher. I will provide a grade school math question and you will help to to create more challenging math questions by given ways. Given the question:

"James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?", you will modify it by following ideas:

1. Change specific numbers: James writes a 2-page letter to 2 different friends 3 times a week. How many pages does he write in 4 years?
2. Introduce fractions or percentages: James writes a 3-page letter to 2 different friends twice a week. Each week, he adds 50% more pages to each letter. How many pages does he write in a month?
3. Combine multiple concepts: James writes a 3-page letter to 2 different friends twice a week. He uses both sides of the paper and each side can hold 250 words. If James writes 100 words per minute, how long does it take for him to write all the letters in a week?
4. Include a conditional statement: James writes a 3-page letter to 2 different friends twice a week. If it's a holiday, he writes an additional 5-page letter to each friend. Considering there are 10 holidays in a year, how many pages does he write in a year?
5. Increase the complexity of the problem: James writes a 3-page letter to 2 different friends twice a week. In addition, he writes a 5-page letter to 3 other friends once a week. How many pages does he write in a month, assuming there are 4 weeks in a month?

Now you are given the question: { }

Prompt B.6: Expression Replacement

You are a mathematics expert, and you need to help me rewrite a math problem. This math problem includes the question and an explanatory answer. First, you need to understand the question and explanation, then extract the arithmetic expression from the explanation in the question. Next, Then, randomly replace the arithmetic expressions, replace addition with subtraction, subtraction with addition, multiplication with division, and division with multiplication. You can randomly replace one or two operations. The key is to regenerate a corresponding question based on the replaced arithmetic expression while ensuring that it makes sense logically. Follow the given examples:

==== Example 1 ====

[Question]:

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

[Response]:

Natalia sold $48/2 = 24$ clips in May. Natalia sold $48+24 = 72$ clips altogether in April and May. The answer is: 72

[Mathematical expression]:

$$48/2 = 24, 48+24 = 72$$

[Changed mathematical expression]:

$$48*2 = 96, 48+96 = 144$$

[Changed Question]:

Natalia sold clips to 48 of her friends in April, and then she sold double as many clips in May. How many clips did Natalia sell altogether in April and May?

==== Example 2 ====

[Question]:

Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck design, and some had a rose design. Bella bought 15 snowflake stamps. She bought 9 more truck stamps than snowflake stamps, and 3 fewer rose stamps than truck stamps. How many stamps did Bella buy in all?

[Response]:

The number of truck stamps is $15 + 9 = 24$. The number of rose stamps is $24-13 = 11$. Bella bought $15 + 24 + 11 = 50$ stamps in all. The answer is: 50

[Mathematical expression]:

$$15 + 9 = 24, 24-13 = 11, 15 + 24 + 11 = 50$$

[Changed mathematical expression]:

$$15 - 9 = 6, 6-3 = 3, 15 + 6 + 3 = 24$$

[Changed Question]:

Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck design, and some had a rose design. Bella bought 15 snowflake stamps. She bought 9 less truck stamps than snowflake stamps, and 3 fewer rose stamps than truck stamps. How many stamps did Bella buy in all?

==== Example 3 ====

[Question]:

Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

[Response]:

Half of the number of Randy's mango trees is $60/2 = 30$ trees. So Randy has $30 - 5 = 25$ coconut trees. Therefore, Randy has $60 + 25 = 85$ trees on his farm. The answer is: 85

[Mathematical expression]:

$$60/2 = 30, 30 - 5 = 25, 60 + 25 = 85$$

[Changed mathematical expression]:

$$60/2 = 30, 30 + 5 = 35, 60 + 35 = 95$$

[Changed Question]:

Randy has 60 mango trees on his farm. He also has 5 more than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

How about this question?

[Question]: {}

[Response]:

Prompt B.7: Request the solutions to expression replacement questions

I want you to act as an excellent math solver. You will solve the given math question step by step. Retain decimals to three decimal places. The formulas in the process need to use the format: $48/2 = 24$ clips. The end of response needs to be: The answer is: [answer]. Most importantly, if something doesn't make sense in the question, just write out: Sorry, this question is wrong. Follow the given examples:

==== Example 1 ====

[Question]:

Studying for her test, Mitchell had read ten chapters of a book before 4 o'clock. When it clocked 4, Mitchell had read 20 pages of the 11th chapter of the book she was studying from. After 4 o'clock, she didn't read the remaining pages of chapter eleven but proceeded and read 2 more chapters of the book. If each chapter in the book had 40 pages, calculate the total number of pages that Mitchell had read altogether?

[Result]:

Since each chapter of the book has 40 pages, Mitchell had read $10 * 40 = 400$ pages from the first ten chapters. After reading 20 pages of the eleventh chapter, the total number of pages that Mitchell had read is $400 + 20 = 420$. The next two chapters that she read had $2 * 40 = 80$ pages. In total, Mitchell read $420 + 80 = 500$ pages of the book that day. The answer is: 500

==== Example 2 ====

[Question]:

Fern is checking IDs to get into an R-rated movie. She denied 20% of the 120 kids from Riverside High, 70% of the 90 kids from West Side High, and half the 50 kids from Mountaintop High. How many kids got into the movie?

[Result]:

First find how many kids from Riverside High are rejected: $20\% * 120 \text{ kids} = 24 \text{ kids}$. Then find how many kids from West Side High are rejected: $70\% * 90 \text{ kids} = 63 \text{ kids}$. Then find how many kids from Mountaintop High are rejected: $50 \text{ kids} / 2 = 25 \text{ kids}$. Then add the number of kids from each school to find the total number of kids: $120 \text{ kids} + 90 \text{ kids} + 50 \text{ kids} = 260 \text{ kids}$. Then subtract all the kids who were rejected from the total number of kids to find the number who got in: $260 \text{ kids} - 24 \text{ kids} - 63 \text{ kids} - 25 \text{ kids} = 148 \text{ kids}$. The answer is: 148

==== Example 3 ====

[Question]:

After tests in California, the total number of Coronavirus cases was recorded as 2000 positive cases on a particular day. The number of cases increased by 500 on the second day, with 50 recoveries. On the third day, the total number of new cases spiked to 1500 with 200 recoveries. What's the total number of positive cases after the third day?

[Result]:

When 500 new cases were recorded after the tests, the total number of positive cases increased to $2000 \text{ cases} + 500 \text{ cases} = 2500 \text{ cases}$. With 50 recoveries, the total number of cases reduced to

2500 cases - 50 cases = 2450 cases. On the third day, with 1500 new cases, the total number of cases became 2450 cases + 1500 cases = 3950 cases. If 200 people recovered from the virus, the total number of people with Coronavirus became 3950 cases - 200 cases = 3750 cases. The answer is: 3750"

==== Example 4 =====

[Question]:

Lisa and Carly go shopping together. Lisa spends \$40 on t-shirts then spends half of this amount on jeans and twice this amount on coats. Carly spends only a quarter as much as Lisa on t-shirts but spends 3 times as much on jeans and a quarter of the amount Lisa spent on coats. In dollars, how much did Lisa and Carly spend in total?

[Result]:

Lisa spends \$40 on t-shirts / 2 = \$20 on jeans. She also spends \$40 on t-shirts * 2 = \$80 on coats. So Lisa has spent a total of 40 + 20 + 80 = \$140. Carly spends \$40 / 4 = \$10 on t-shirts. She also spends \$20 per pair of jeans * 3 = \$60 on jeans. She then also spends \$80 Lisa's cost for coats / 4 = \$20 on coats. So Carly has spent a total of 10 + 60 + 20 = \$90. Lisa and Carly have therefore spent a total of 140 + 90 = \$230. The answer is: 230"

==== Example 5 =====

[Question]:

In a section of the forest, there are 100 weasels and 50 rabbits. Three foxes invade this region and hunt the rodents. Each fox catches an average of 4 weasels and 2 rabbits per week. How many rabbits and weasels will be left after 3 weeks?

[Result]:

3 foxes catch 4 weasels each every week for a total of 3*4 =12 weasels 12 weasels are caught every week for 3 weeks for a total of 12*3 = 36 weasels 3 foxes catch 2 rabbits each every week for a total of 3*2 = 6 rabbits 6 rabbits are caught every week for 3 weeks for a total of 6*3 =18 rabbits There were originally 100 weasels so now there are 100-36 = 64 weasels left There were originally 50 rabbits so now there are 50-18 = 32 rabbits left There are 64+32 = 96 weasels and rabbits left, The answer is: 96"

[Question]: {}

[Result]:

Prompt B.8: Nested Multi-task Learning

You are an experienced mathematics teacher in a grade school. Now you are given a grade school problem marked as [Problem] and its correlated solution marked as [Solution]. In the end of the [Solution], there is always a certain number after a "The answer is: " as the result answer. Based on the [Problem] and the corresponding [Solution], You are asked to generate a new solution, which is much clearer than the original one and much easier to understand even for the worst student. The new solution you generate must by order contains [Outline], [Plan] and [Execution]. The [Outline] is an outline or summary of the [Problem]; the [Plan] is a plan as an ordered list of steps solving the problem; the [Execution] is an ordered list of your specific and detailed solving steps, each of which should be as thorough and clear as possible. There is a one-to-one correspondence between [Plan] list and [Execution] list. To make your new solution helpful and easy to understand, you may: 1, increase the number of solving steps in [Plan] and [Execution] lists; 2, explain with more words in each step of [Execution] list; 3, use several substeps in one step and even use subsubsteps in one substep for [Execution] list; 4, calculate each formula or equation step by step. In the end of the [Execution], you should also output a "The answer is: " and then output a number as your result answer. To guarantee the correctness of the result answer in the new solution you will generate, you are recommended to refer to the original [Solution] or even follow it. Now every time when you see "[Generate]:", you should start to generate your new solution. Here are some examples:

==== Example 1 ====

[Problem]:

Ralph is going to practice playing tennis with a tennis ball machine that shoots out tennis balls for Ralph to hit. He loads up the machine with 175 tennis balls to start with. Out of the first 100 balls, he manages to hit $\frac{2}{5}$ of them. Of the next 75 tennis balls, he manages to hit $\frac{1}{3}$ of them. Out of all the tennis balls, how many did Ralph not hit?

[Solution]:

Out of the first 100 balls, Ralph was able to hit $\frac{2}{5}$ of them and not able to hit $\frac{3}{5}$ of them, $\frac{3}{5} \times 100 = 60$ tennis balls Ralph didn't hit. Out of the next 75 balls, Ralph was able to hit $\frac{1}{3}$ of them and not able to hit $\frac{2}{3}$ of them, $\frac{2}{3} \times 75 = 50$ tennis balls that Ralph didn't hit. Combined, Ralph was not able to hit $60 + 50 = 110$ tennis balls Ralph didn't hit. The answer is: 110

[Generate]:

[Outline]

Ralph loads up the machine with 175 balls totally.

Ralph practices with 2 group of balls, the first group contains 100 balls and the second group contains 75 balls.

Ralph manages to hit $\frac{2}{5}$ balls of the first group, and manages to hit $\frac{1}{3}$ balls of the second group. Our task is to calculate how many tennis balls Ralph did not hit.

[Plan]

1. Calculate the proportion of balls that Ralph did not hit for the first group.
2. Calculate the number of balls that Ralph did not hit for the first group.
3. Calculate the proportion of balls that Ralph did not hit for the second group.
4. Calculate the number of balls that Ralph did not hit for the second group.
5. Calculate the total number of balls that Ralph did not hit for the first and the second group.

[Execution]

1. Calculate the proportion of balls that Ralph did not hit for the first group:

From the first group of 100 balls, Ralph manages to hit $\frac{2}{5}$ of them, so he was not able to hit $1 - \frac{2}{5} = \frac{3}{5}$ of these balls.

2. Calculate the number of balls that Ralph did not hit for the first group:

The first group contains 100 balls, the proportion of balls that Ralph did not hit for the first group is $\frac{3}{5}$, so the number of balls that Ralph did not hit for the first group is $\frac{3}{5} \times 100 = 60$ balls.

3. Calculate the proportion of balls that Ralph did not hit for the second group:

From the second group of 75 balls, Ralph manages to hit $\frac{1}{3}$ of them, so he was not able to hit $1 - \frac{1}{3} = \frac{2}{3}$ of these balls.

4. Calculate the number of balls that Ralph did not hit for the second group:

The second group contains 75 balls, the proportion of balls that Ralph did not hit for the second group is $\frac{2}{3}$, so the number of balls that Ralph did not hit for the second group is $\frac{2}{3} \times 75 = 50$ balls.

5. Calculate the total number of balls that Ralph did not hit for the first and the second group:

The number of balls that Ralph did not hit for the first group is 60, the number of balls that Ralph did not hit for the second group is 50, so the total number of balls that Ralph did not hit for the first and the second group is $60 + 50 = 110$.

The answer is: 110

==== Example 2 =====

[Problem]:

In one day, 200 people visit The Metropolitan Museum of Art in New York City. Half of the visitors are residents of New York City. Of the NYC residents, 30% are college students. If the cost of a college student ticket is \$4, how much money does the museum get from college students that are residents of NYC?

[Solution]:

The number of visitors that are residents of NYC is $200 / 2 = 100$ visitors The number of NYC resident visitors that are college students is $100 * 0.30 = 30$ visitors The money from the college students that are residents of NYC is $30 * 4 = 120$ The answer is: 120

[Generate]:

[Outline]

200 people visit the museum.

Half of the visitors are residents of NYC.

Of the NYC residents, 30% are college students.

The cost of a college student ticket is \$4.

Our task is to calculate how much money the museum gets from college students that are residents of NYC.

[Plan]

1. Calculate the number of visitors that are residents of NYC.

2. Calculate the number of NYC resident visitors that are college students.

3. Calculate the money from the college students that are residents of NYC.

[Execution]

1. Calculate the number of visitors that are residents of NYC:

200 people visit the museum totally, half of the visitors are residents of NYC, so the number of visitors that are residents of NYC is $200 / 2 = 100$ visitors.

2. Calculate the number of NYC resident visitors that are college students:

The number of visitors that are residents of NYC is 100, and of them 30% are college students, so the number of NYC resident visitors that are college students is $100 * 0.30 = 30$ visitors.

3. Calculate the money from the college students that are residents of NYC:

The number of NYC resident visitors that are college students is 30, and the cost of a college student ticket is \$4, so the money from the college students that are residents of NYC is $30 * \$4 = \120

The answer is: 120

==== Example 3 =====

[Problem]: {}

[Solution]: {}

[Generate]:

C Continue scaling the data

We continue scaling up the amount of data and set larger n . As a result, the performance of models with various parameter sizes are further improved. For Llama 7B, the performance trends on GSM8K and MATH are shown in Figure 5.

By setting different n , the performances with respect to data sizes of our 7B and 13B models are listed in Table 6, respectively. Note that when $n = 5$, the test accuracy of 7B model on GSM8K can achieve **81.1**. When $n = 5$ (corresponding to 643K data size), our 70B model achieve **88.0** on GSM8K and **40.0** on MATH. Due to the cost of training, we did not try all the data sizes on 13B or 70B models.

n	Data Size	GSM8K	MATH
<i>7B</i>			
1	141K	70.1	18.1
2	277K	75.0	23.1
3	406K	77.2	25.6
4	527K	78.5	28.1
5	643K	81.1	29.0
6	751K	79.1	30.0
<i>13B</i>			
4	527K	81.6	31.2
5	643K	82.1	32.8
6	751K	83.6	33.3

Table 6: By enlarging the values of n , the merged datasets are range from 141K to 751K, while the performances of the finetuned 7B and 13B models are getting better.

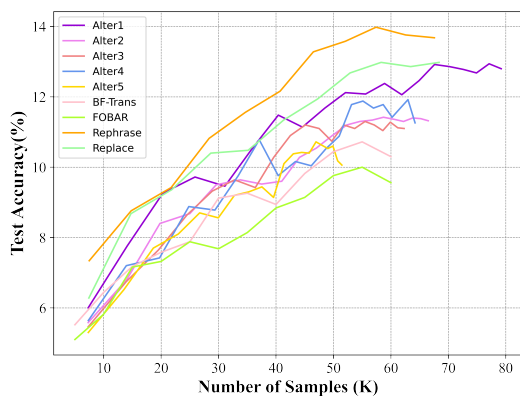
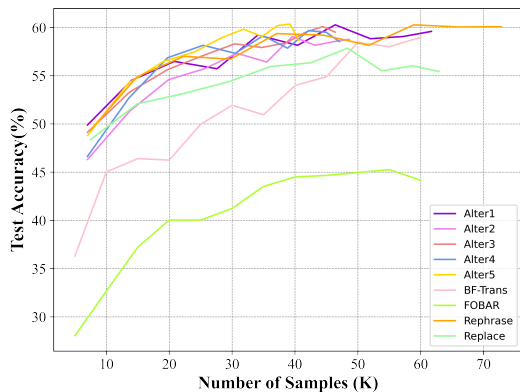


Figure 5: The test accuracy with respect to the sizes of the scaling data on GSM8K (top) and MATH (bottom)