



# JiuZhang3.0: Efficiently Improving Mathematical Reasoning by Training Small Data Synthesis Models

Kun Zhou<sup>1\*</sup>, Beichen Zhang<sup>2\*</sup>, Jiapeng Wang<sup>2</sup>, Zhipeng Chen<sup>2</sup>, Wayne Xin Zhao<sup>2†</sup>,  
Jing Sha<sup>3</sup>, Zhichao Sheng<sup>3</sup>, Shijin Wang<sup>3,4†</sup>, Ji-Rong Wen<sup>2</sup>

<sup>1</sup>School of Information, Renmin University of China

<sup>2</sup>Gaoling School of Artificial Intelligence, Renmin University of China

<sup>3</sup>iFLYTEK Research <sup>4</sup>iFLYTEK AI Research (Central China)

francis\_kun\_zhou@163.com, {zhangbeichen724, batmanfly}@gmail.com,  
{wangjip1010, zhipeng\_chen, jrwen}@ruc.edu.cn, {jingsha, sjwang3}@iflytek.com

## Abstract

Mathematical reasoning is an important capability of large language models (LLMs) for real-world applications. To enhance this capability, existing work either collects large-scale math-related texts for pre-training, or relies on stronger LLMs (*e.g.*, GPT-4) to synthesize massive math problems. Both types of work generally lead to large costs in training or synthesis. To reduce the cost, based on open-source available texts, we propose an efficient way that trains a small LLM for math problem synthesis, to efficiently generate sufficient high-quality pre-training data. To achieve it, we create a dataset using GPT-4 to distill its data synthesis capability into the small LLM. Concretely, we craft a set of prompts based on human education stages to guide GPT-4, to synthesize problems covering diverse math knowledge and difficulty levels. Besides, we adopt the gradient-based influence estimation method to select the most valuable math-related texts. The both are fed into GPT-4 for creating the knowledge distillation dataset to train the small LLM. We leverage it to synthesize 6 million math problems for pre-training our JiuZhang3.0 model. The whole process only needs to invoke GPT-4 API 9.3k times and use 4.6B data for training. Experimental results have shown that JiuZhang3.0 achieves state-of-the-art performance on several mathematical reasoning datasets, under both natural language reasoning and tool manipulation settings. Our code and data will be publicly released in <https://github.com/RUCAIBox/JiuZhang3.0>.

## 1 Introduction

Large language models (LLMs) have shown remarkable capabilities on a variety of tasks [1–3]. However, they still struggle in solving complex mathematical problems [4]. Recent work has shown that it is an effective approach to training LLMs on math-related data for improving the mathematical reasoning ability [5, 6]. Typically, they either collect the math-related data from the available corpora (*e.g.*, webpages and books) for pre-training [7–9], or rely on stronger LLMs to synthesize high-quality math problems for fine-tuning [10–12]. Despite the success, existing approaches would generally cause large training or inference costs. Due to the complexity and diversity of mathematical problems, the former type of work mostly needs to collect a large-scale corpus (*e.g.*, 120B data for Deepseek-Math) for training, which greatly increases the training cost [5, 7, 8]. Similarly, to guarantee the knowledge coverage and effectiveness of the synthetic problems, the latter type of work relies on

\*Equal contributions.

†Corresponding authors.

stronger LLMs with larger scales (*e.g.*, GPT-4) to create massive math problems, leading to larger inference cost [6, 10, 11]. In Figure 1, we show our estimated total costs of re-implementing two math-related LLMs (>\$40000), details are in Appendix A

In this work, we aim to develop a relatively low-cost data synthesis approach for improving the mathematical reasoning abilities of LLMs. Our key idea is that **the data synthesis capability can be well learned by small LLMs**. Here, *small* is a relative wording, which is in contrast with the extremely large or costly data synthesis models used in prior studies [6, 13], such as GPT-4 or Qwen-72B. Actually, existing work [10, 14, 15] has extensive evidence of strong learning and adaptation abilities of small LLMs for new tasks and domains with suitable strategies (*e.g.*, training with high-quality supervised data), including math, science, and complex multimodal tasks. However, this exploration has been neglected in prior efforts on data synthesis. This attempt can be essentially generalized to a broader problem: *whether a small (or weak) model can produce high-quality data that is useful for training a large (or strong) model?* Inspired by this motivation, we seek to train a relatively small yet powerful LLM for synthesizing high-quality math-related data.

However, due to the diverse and complex nature of math problems, it is challenging to train well-performing small LLMs for synthesizing high-quality ones. Although we can leverage GPT-4, it is not efficient to use it for synthesizing a large-scale knowledge distillation (KD) dataset. Specifically, we aim to build the KD dataset through a low-cost strategy, but it can sufficiently capture diverse and useful knowledge about math problem synthesis. Thus, we should guarantee the knowledge coverage and usefulness of the instances within the KD dataset. To achieve it, we first craft a set of prompts, and each prompt corresponds to an education stage of humans, *e.g.*, middle school and college. Using the above prompts, the instances within the dataset can well cover broad mathematical knowledge and different difficulty levels. Besides, for usefulness, we estimate the influence of the available math-related texts and downstream tasks, by computing the gradient similarity between their corresponding synthetic data and the task instances. Then, we select the top-ranking math-related texts into the KD dataset, which are high-value ones with more positive influence on downstream tasks. By feeding the texts with prompts into GPT-4, we collect the outputs to build the KD dataset.

Based on the KD dataset, we train DeepSeekMath-7B as our data synthesis model, which is much smaller than other commonly-used LLMs in existing work [6, 12, 13], *e.g.*, GPT-4 and Qwen-72B. Owing to our data selection strategy, we only require GPT-4 to generate 9,335 instances based on the selected most valuable texts for training it. Then, we utilize the crafted prompts to guide it for synthesizing high-quality problems. Benefiting from the strong data synthesis capability of the small model, we only need to synthesize 5,984,584 high-quality math problems (4.6B tokens) for pre-training our JiuZhang3.0. Thus, the total inference and training cost is much less than existing work, as shown in Figure 1. After pre-training, we also collect open-source math instructions to fine-tune JiuZhang3.0. The experimental results have shown that JiuZhang3.0 can mostly outperform state-of-the-art methods across 18 evaluation datasets, in both the natural language reasoning and tool manipulation settings. Our contributions are summarized as follows:

- (1) our research provides compelling evidence that it is feasible to efficiently train a small LLM (7B) for synthesizing training data to improve the mathematical reasoning of LLMs. As results shown in Section 4.3, its synthetic data is more useful than larger LLMs in improving the performance.
- (2) we propose an efficient solution for training LLMs to improve mathematical reasoning, which only needs to invoke GPT-4 API 9.3k times and pre-train on 4.6B high-quality synthetic data, with nearly 20% total cost of existing state-of-the-art methods.
- (3) JiuZhang3.0 achieves state-of-the-art performance among open-source LLMs on several tasks and settings, *e.g.*, 52.8 (JiuZhang3.0-7B) vs. 50.2 (DeepSeekMath-7B-RL) on MATH, 89.8 (JiuZhang3.0-8×7B) vs. 86.4 (MAmmoTH2-8×7B-Plus) on GSM8k in the natural language reasoning setting.

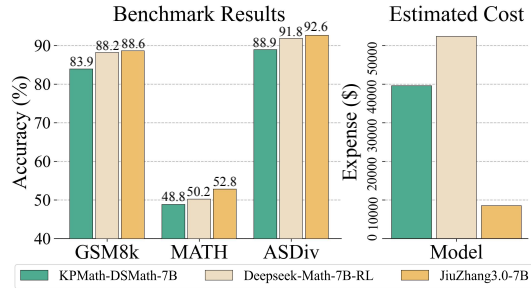


Figure 1: The comparison of existing work and our method in task performance and the total cost.

## 2 Related Work

**Large Language Models.** LLMs have demonstrated remarkable capabilities in a variety of NLP tasks, and commercial LLMs like ChatGPT, Claude, and Gemini [2, 16, 17], represent cutting-edge capabilities. Meanwhile, the performance of open-source models (*e.g.*, LLaMA-3, Mixtral) has also developed rapidly [18, 19]. To further improve the capability of LLMs on special tasks or domains, existing work mainly focuses on optimizing the following aspects: (1) prompt engineering such as chain-of-thought and tree-of-thought [20, 21]; (2) continual pre-training on a domain-specific or task-specific corpus, improving the model to deal with downstream tasks [5, 7, 8, 22, 23]; (3) supervised fine-tuning, which involves fine-tuning the model on related instruction datasets, enhancing LLMs to follow special task instructions [24, 25]; (4) other strategies including RLHF [26], tool augmentation [27], decoding optimization [28] and *et al.* We aim to efficiently improve the capability of LLMs for mathematical reasoning by pre-training on synthetic data.

**Mathematical Reasoning.** Despite the impressive progress, mathematical reasoning remains a weak aspect of LLMs. To enhance LLMs’ ability in mathematical reasoning, researchers have proposed a surge of methods from the aspects of prompting, pre-training and fine-tuning. For prompting, the chain-of-thought (CoT) prompts have been widely used to guide LLMs for performing multi-step reasoning on complex math problems [20]. Based on it, following work utilizes tools [27, 29–33] and verifiers [9, 34–36], to further improve the accuracy of the mathematical reasoning process. For pre-training, existing work [5, 7, 9, 37, 38] collects a large-scale math-related corpus and continually pre-training open-source LLMs on it. Supervised fine-tuning methods focus on using relatively less high-quality data for training the LLM, which are typically math-related instructions [10, 12, 39]. Recent studies show that the complexity of mathematical reasoning demands high-quality instruction pairs, leading to reliance on advanced LLMs like GPT-4 for data synthesis [6, 40, 41]. The pre-training and fine-tuning methods generally lead to large training and data annotation costs, respectively. Our work aims to train a small LLM specially for math problem synthesis, which can efficiently produce sufficient data for training.

**Data Synthesis.** For complex tasks and scenarios (*e.g.*, mathematical reasoning), it is necessary to collect a substantial amount of data for training the LLM to enhance it. However, the available data may not be sufficient, hence researchers have explored using automatically synthetic data with consistent distribution to real data, to enrich the training corpus [42–48]. For data synthesis on mathematical reasoning tasks, existing work can be roughly categorized into the following two types, according to their based guided information. The first type of work starts with existing problems or math-related texts to synthesize similar problems or solutions [10, 40, 41, 49, 50]. The other type of work relies on available knowledge points, and devises special prompts to guide LLMs for synthesizing related problems with the solutions [6, 51]. As correctness is important, the two types of work generally design rules to check and remove wrong ones. In this work, based on the data synthesis model, we also construct the multi-source math corpora, and craft several prompts to guide it in producing diverse and useful math problems.

## 3 Approach

In this section, we present our approach that aims to train a small LLM for synthesizing math problems. First, we initialize the data synthesis model by training it on the KD dataset, composed of crafted prompts, randomly sampled math-related texts, and the corresponding synthetic problems and solutions from GPT-4. Then, we improve its data synthesis capability by retraining it on the updated knowledge distillation dataset, where we add the high-value math-related texts selected by gradient-based influence estimation strategy. Finally, we utilize the model to synthesize massive high-quality math problems for training JiuZhang3.0, based on the multi-source math-related corpus.

### 3.1 Preliminary

We focus on training a small data synthesis LLM, for synthesizing high-quality math problem-solution pairs to pre-train LLMs and improve its mathematical reasoning capability. To guarantee the quality of the synthetic data, we utilize GPT-4 to create the knowledge distillation (KD) dataset  $\mathcal{D}_{KD} = \{p_i, t_i, \hat{q}_i, \hat{s}_i\}_{i=1}^N$  for training the small LLM, where the math-related text  $t_i$  and the prompt

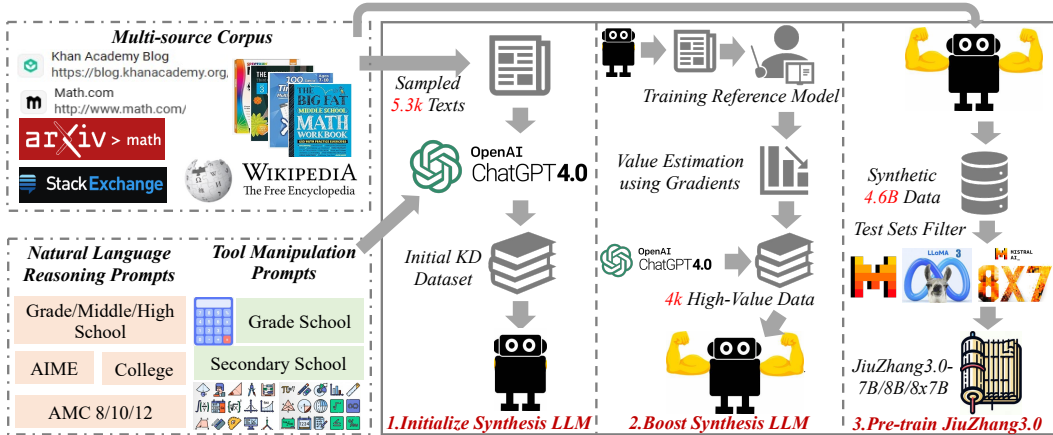


Figure 2: The pipeline of our approach. We first initialize the data synthesis LLM by distilling the knowledge from GPT-4 on randomly sampled data, then boost it using the high-value data selected by gradient-based value estimation strategy, finally utilize it for synthesizing data to train JiuZhang3.0.

$p_i$  are the input of GPT-4,  $\hat{q}_i$  and  $\hat{s}_i$  are its synthetic math problem and solution respectively. Then, we train the small LLM on  $\mathcal{D}_{KD}$  to imitate the data synthesis ability of GPT-4. Finally, we leverage it to synthesize the pre-training dataset  $\mathcal{D} = \{q_i, s_i\}_{i=1}^M$  based on all the collected math-related texts with randomly selected prompts  $\{p_i, t_i\}_{i=1}^M$ , which are used for training our JiuZhang3.0.

## 3.2 Initializing Data Synthesis Model

In this work, we consider the natural language reasoning and tool manipulation settings, where LLMs require solving the problem by generating a natural language solution [20] and an executable program with external interpreters [52], respectively. Thus, we train the data synthesis LLMs on the initial KD dataset, containing special prompts, math-related texts, and GPT-4 outputs for the two settings.

### 3.2.1 Prompts for Math Problem Synthesis

We aim to craft a prompt set that can well cover the knowledge points and difficulty levels in human math education. Thus, for the natural language reasoning and tool manipulation settings, we manually craft prompt templates respectively, and each corresponds to a certain education stage.

**Prompts for Natural Language Reasoning.** We consider the following 4 human education stages and 4 worldwide competitions, *i.e.*, Grade School, Middle School, High School, and College; AMC (American Mathematics Competition) 8, AMC 10, AMC 12, and AIME (American Invitational Mathematics Examination). Based on these, we design 8 prompts with corresponding instructions and guidelines. We show an example of grade school math problem synthesis:

**## Instruction:** Create an age-appropriate math word problem for grade school students based on the provided math content.

**## Guidelines:** [Problem]: Craft a concise math word problem suitable for grade school, focusing on basic arithmetic operations, number sense, simple shapes, ... [Solution]: Provide a clear, step-by-step solution to the problem using simple language that a grade school student could understand, ...

**Prompts for Tool Manipulation.** We consider 2 types of math problems, *i.e.*, Grade School and Secondary School Competitions, as the competition math problems may need tools for advanced math operations (*e.g.*, integral computation), while grade school math problems are much easier and can be solved by basic operations (*e.g.*,  $+*$ ). As tool manipulation solves the problem via executable programs, we use more words in prompts to emphasize the data format and show an example as follows:

**## Instruction:** Please gain inspiration from the following random math content to create a high-quality ... Present your output in two distinct sections: [Problem Description] and [Solution].

**## Guidelines:** [Problem]: This should be completely self-contained, providing all the contextual information one needs to understand, ... [Solution]: Offer a comprehensive, correct solution that accurately addresses the [Problem] you provided using Python code, ...

**## Example:**[Problem Description] Janet buys 3 pounds of broccoli, ... [/Problem Description] [Solution] def spending(): cost = 4 ...[/Solution]

### 3.2.2 Knowledge Distillation from GPT-4

Based on the prompts, we build the KD dataset to train our data synthesis LLM. We randomly sample 5,336 math-related texts, and concatenate each one with a randomly selected prompt, to compose the input *i.e.*,  $[p_i; w_i]$ . Then, we feed it into GPT-4, and extract the synthetic math problem  $\hat{q}_i$  and solution  $\hat{s}_i$  from its output using regular expressions, to compose the KD dataset  $\mathcal{D}_{KD} = \{p_i, w_i, \hat{q}_i, \hat{s}_i\}_{i=1}^{N_{ini}}$ . Next, we utilize it to train the synthesis model, and the learning objective is:

$$L(\theta_{syn}) = \sum_{i=1}^{N_{ini}} \log P([\hat{q}_i; \hat{s}_i] | [p_i; w_i]), \quad (1)$$

where  $\theta_{syn}$  denotes the parameters of the data synthesis LLM. In this way, we can teach it to generate new math problem-solution pairs based on prompts and math-related texts by imitating GPT-4.

### 3.3 Boosting Synthesis Model using High-Value Data

After initialization, we further improve the data synthesis LLM using high-value KD data. However, it would be costly if we first utilize GPT-4 to generate candidates and then select high-value ones. Therefore, we propose an efficient way that leverages the data synthesis LLM for generating candidates, and then selects valuable ones to feed into GPT-4. Specifically, we incorporate the gradient-based method [53] to estimate the influence of each synthetic instance for downstream math-related tasks, and select the top-ranking ones to update the KD dataset for retraining the data synthesis LLM.

#### 3.3.1 Gradient-based Data Value Estimation

According to the influence formulation [54], at a certain training step of a model parameterized by  $\theta$ , the influence of a training instance  $z$  on another instance  $z'$  can be estimated by computing the similarity between their produced gradients, denoted as:

$$\text{Inf}(z, z') \propto \text{Sim}(\nabla l(z, \theta), \nabla l(z', \theta)). \quad (2)$$

By using it, we can measure the value of each synthetic data by computing its gradient similarity with downstream math-related task data. Concretely, we first train a reference model using LoRA, then compute its gradients on LoRA parameters as the features, to help estimate the data value.

**Training Reference Model using LoRA.** Inspired by existing work [53], we train a LLM for mathematical problem solving using LoRA [55] as the reference model. As LoRA only requires to optimize the low-rank adapters in the LLM, we can efficiently train the reference model on limited computation resources, and reduce the number of trainable parameters for efficient computation of gradient similarity. Besides, to further reduce the training cost, we randomly select a subset of synthetic math problems generated by the data synthesis model, denoted as  $\mathcal{D}_{lora} = \{q_i, s_i\}_{i=1}^{M_l}$ . Then, we train the reference model to predict the solution based on the given problem, denoted as:

$$L_{ref}(\theta_{lora}) = \sum_{i=1}^{M_l} \log P(s_i | q_i) \quad (3)$$

where  $\theta_{lora}$  denotes the parameters of LoRA, and  $M_l$  is the number of training data.

**Computing Gradient Features.** After training the reference model, we compute the gradients of LoRA parameters as the feature of each synthetic instance. As their dimension is large, we follow existing work [56] that performs random projection to obtain the low-dimensional features as:

$$\hat{\nabla} l_{ref}(z, \theta_{lora}) = \Pi^\top \nabla l_{ref}(z, \theta_{lora}), \quad (4)$$

where  $z = \langle p_i, w_i, q_i, s_i \rangle$  denotes a synthetic instance,  $\Pi \in \mathbb{R}^{d' \times d}$  is a projection matrix initialized by the Rademacher distribution, its entries are -1 or 1,  $d'$  and  $d$  are the dimensions before and after projection, respectively. According to the Johnson-Lindenstrauss Lemmas [57], this operation can nearly preserve the gradient distances, ensuring the usefulness of the low-dimensional features.

**Estimating Data Value.** By using Eq. 4, we can compute the gradient features for synthetic instances. Then, we randomly sample  $M_D$  instances from the training sets of downstream math-related datasets  $\{z'_i\}_{i=1}^{M_D}$ , where  $z'_i = \langle \tilde{q}_i, \tilde{s}_i \rangle$ , and also compute their gradient features. Next, we estimate the value of each synthetic instance by computing the similarity between its gradient feature and the average feature of all the sampled downstream instances as:

$$V(z) = \text{Cosine}(\hat{\nabla}l_{ref}(z, \theta_{lora}), \frac{1}{M_D} \sum_{i=1}^{M_D} \hat{\nabla}l_{ref}(z'_i, \theta_{lora})), \quad (5)$$

where  $\text{Cosine}(x, y)$  computes the cosine similarity between the two vectors. In this way, the instance with higher data value would lead to a more positive influence on the downstream math-related tasks.

### 3.3.2 Retraining Data Synthesis Model

Based on the estimated values, we can rank all the synthetic instances, and the top-ranking  $N_{add}$  ones can be regarded as the most valuable data  $\{\langle p_i, w_i, q_i, s_i \rangle\}_{i=1}^{N_{add}}$  for improving downstream math-related tasks. Thus, we utilize GPT-4 to regenerate the synthetic math problems based on their prompts and original math-related texts, to acquire corresponding more high-quality math problems and solutions. Then, we add the new GPT-4 synthetic data  $\{p_i, w_i, \hat{q}_i, \hat{s}_i\}_{i=1}^{N_{add}}$  into the KD dataset, and the new data is capable of guiding the small LLM to generate more useful math problems for downstream tasks. Next, we retrain the data synthesis LLM with the updated KD dataset using Eq. 1.

## 3.4 Pre-training JiuZhang3.0 using Synthetic Data

After training the data synthesis LLM, we construct the multi-source corpus containing rich math-related texts to cover more knowledge and scenarios. Then, we synthesize massive math problems based on it, which are used for pre-training JiuZhang3.0.

**Constructing Multi-source Corpus.** We consider the following data types and select the corresponding open-source datasets to compose the math-related multi-source corpus.

- *Webpages*: we use the OpenWebText corpus [58], which consists of 6.3M math-related web documents extracted from Common Crawl.
- *Books*: we use the Mathpile-textbook dataset [59], including 4K educational textbooks, lecture notes and synthetic books.
- *Papers*: we use the Mathpile-Arxiv dataset [59], and select the high-quality ones according to the estimated scores (0.6-0.9), which are released by AutoMathText [60].
- *QA Data*: we select the StackExchange subset of the MMIQC dataset [41], which contains 1.2M processed real-world math question-answering pairs.
- *Wikipedia*: we use the Mathpile-Wikipedia dataset [59], consisting of 106K documents from math-related entries in Wikipedia.

**Data Synthesis for Training JiuZhang3.0.** For each instance within the multi-source corpus, we randomly select a prompt from the prompt set and embed the text into the prompt to compose the input. Then, we feed inputs into the data synthesis model, to generate the math problems and solutions for composing the synthesis dataset  $\mathcal{D} = \{q_i, s_i\}_{i=1}^M$ . Here, we follow existing work [5, 13] to filter out the instances with 10-grams overlap to both inputs and outputs from test sets of downstream evaluation tasks. We synthesize about 6M math problems (4.6B tokens) in total, which are used for pre-training JiuZhang3.0 to predict the solution based on the given problems.

Table 1: Results on 6 datasets in the natural language reasoning setting. The best and second-best ones among LLMs with similar scales are marked in bold and underlined respectively.

Models	GSM8k	MATH	SVAMP	ASDiv	MAWPS	CARP	Avg.
ChatGPT	76.6	38.2	83.7	87.7	96.9	41.3	70.7
GPT-4	92.2	65.4	92.9	94.3	96.6	53.6	82.5
Qwen-1.5-110B	85.4	<u>49.4</u>	86.2	85.1	94.3	<b>53.6</b>	75.7
Qwen-1.5-72B	77.6	39.4	83.1	85.1	95.8	<u>53.0</u>	72.3
Mixtral-8×7B	74.4	29.0	76.5	78.5	93.9	38.8	65.2
Llemma-34B	60.2	24.6	68.0	75.6	89.8	36.5	59.1
Intern-Math-20B	64.9	27.4	74.9	79.6	94.4	42.3	63.9
ChatGLM-Math-32B	82.6	40.6	-	-	-	-	-
MAmmoTH2-8x7B-Plus	<u>86.4</u>	47.0	<u>90.0</u>	<u>92.2</u>	<b>97.0</b>	45.8	<u>76.4</u>
JiuZhang3.0-8x7B (Ours)	<b>89.8</b>	<b>53.8</b>	<b>90.2</b>	<b>93.1</b>	<u>96.7</u>	52.3	<b>79.3</b>
DeepSeek-7B	13.6	4.8	40.8	52.1	65.4	10.3	31.2
Mistral-7B	41.2	13.6	64.7	68.5	87.5	14.9	48.4
LLaMA-3-8B	54.5	19.6	68.5	72.8	90.5	29.2	55.9
Gemma-7B	54.1	19.6	69.7	74.2	89.0	30.5	56.2
Qwen-1.5-7B	60.5	28.2	64.9	74.9	90.1	38.6	59.5
Llemma-7B	39.2	18.4	56.9	69.0	82.7	31.8	49.7
InternLM-Math-7B	45.9	15.8	67.3	71.2	88.3	28.0	52.8
Rho-1-Math-7B	66.3	31.0	78.5	79.2	94.0	36.7	64.3
DeepSeekMath-7B	64.1	34.2	73.7	82.7	92.7	44.4	65.3
Mistral-7B-MMIQC	75.0	34.2	73.5	82.1	90.1	36.5	65.2
MetaMath-Mistral-7B	77.8	29.6	79.6	81.2	93.7	30.5	65.4
Abel-7B-002	80.4	29.6	78.8	82.7	93.5	33.2	66.4
WizardMath-7B-1.1	82.2	32.8	80.7	84.2	93.8	31.9	67.6
Math-Shepherd-Mistral-7B	84.3	34.4	82.9	82.8	92.5	32.9	68.3
KPMath-DSMath-7B	83.9	48.8	81.5	88.9	94.8	-	-
MAmmoTH2-7B-Plus	84.2	46.2	<u>90.3</u>	90.3	95.8	44.3	75.2
MAmmoTH2-8B-Plus	84.4	41.2	89.9	89.9	<u>97.1</u>	44.8	74.6
DeepSeekMath-7B-Instruct	82.3	45.8	83.7	90.1	95.7	45.8	73.9
DeepSeekMath-7B-RL	88.2	50.2	87.3	91.8	95.5	<b>51.6</b>	77.4
JiuZhang3.0-7B (Ours)	<b>88.6</b>	<b>52.8</b>	<b>90.4</b>	<b>92.6</b>	<b>97.3</b>	<u>51.0</u>	<b>78.8</b>
JiuZhang3.0-8B (Ours)	<b>88.6</b>	<u>51.0</u>	89.4	<b>92.6</b>	<u>97.1</u>	50.9	<u>78.3</u>

## 4 Experiments

### 4.1 Experimental Settings

For our JiuZhang3.0, we follow existing work [13] that train the 7B, 8B and 8×7B versions based on Mistral-7B [61], LLaMA-3-8B [18], and Mixtral-8×7B [19]. During training, we first pre-train it on our synthetic 4.6B math problem-solution pairs and then fine-tune it on the collected multiple open-source instruction datasets. We evaluate JiuZhang3.0 in two settings, *i.e.*, natural language reasoning and tool manipulation. More details about the fine-tuning data, evaluation datasets, baseline methods, and implementation details are in Appendix B, C, D and E, respectively.

### 4.2 Results and Analysis

**Natural Language Reasoning.** The results of this setting are shown in Table 1. First, the baseline methods trained on math-related data perform better than others. Among them, DeepSeekMath-7B is the best-performed base LLM, and DeepSeekMath-7B-RL also performs better than other baselines, since they have been pre-trained on 120B corpus containing rich math-related data. Besides, KPMath-DSMath-7B and MAmmoTH2 also perform well. Concretely, KPMath-DSMath-7B is trained on nearly 1M synthetic math problems produced by GPT-4, and MAmmoTH2 also utilizes the GPT-4, Mixtral-8×7B, and Qwen-72B to extract and refine the problems existing in the webpages. The acquired problems can greatly improve their performance in math problem solving. In our approach, we also utilize synthetic math problems to train our JiuZhang3.0-7B and 8B models. Differently,

Table 2: Results on 5 other datasets with different data formats or related to interdisciplinary fields, and we abbreviate MMLU-STEM into M-STEM. The best and second-best methods among LLMs with similar scales are marked in bold and underlined respectively.

Models	TabMWP	AQuA	SAT-Math	M-STEM	OCW-Math	Avg.
ChatGPT	82.0	53.9	78.1	63.5	11.0	57.7
GPT-4	90.8	76.9	96.9	77.1	26.5	73.6
Qwen-1.5-110B	80.5	64.6	<b>87.5</b>	71.5	14.0	63.6
Qwen-1.5-72B	56.1	55.1	<b>87.5</b>	68.8	7.7	55.0
Mixtral-8x7B	67.3	48.0	65.6	62.3	8.8	50.4
Llemma-34B	57.1	46.1	71.9	54.3	11.8	48.2
Intern-Math-20B	63.4	44.1	65.6	62.3	7.0	48.5
MAmmoTH2-8x7B-Plus	62.7	55.9	81.2	<b>71.8</b>	<u>18.8</u>	58.1
JiuZhang3.0-8x7B (Ours)	<b>84.7</b>	<b>65.4</b>	81.2	66.9	<b>23.5</b>	<b>64.3</b>
Mistral-7B	37.3	34.3	56.2	49.5	3.3	36.1
LLaMA-3-8B	67.5	46.5	56.2	54.4	7.7	46.5
Gemma-7B	60.9	42.9	71.9	57.7	4.8	47.6
Llemma-7B	49.2	37.8	62.5	45.8	7.7	40.6
Rho-1-Math-7B	55.5	49.2	75.0	54.9	6.2	48.2
DeepSeekMath-7B	69.8	51.6	<u>84.4</u>	56.1	17.6	55.9
DeepSeekMath-7B-Instruct	70.5	60.6	<u>84.4</u>	57.9	19.5	58.6
MAmmoTH2-7B-Plus	54.7	<b>62.2</b>	<u>84.4</u>	64.0	15.1	56.1
MAmmoTH2-8B-Plus	<u>75.1</u>	57.5	<b>87.5</b>	<b>65.7</b>	14.7	<u>60.1</u>
JiuZhang3.0-7B (Ours)	74.8	59.4	81.2	53.6	<u>20.2</u>	57.8
JiuZhang3.0-8B (Ours)	<b>79.2</b>	<b>62.2</b>	<u>84.4</u>	60.4	<b>21.3</b>	<b>61.5</b>

Table 3: Results on 6 mathematical reasoning datasets under the tool manipulation setting. The best and second-best methods are marked in bold and underlined respectively.

Models	GSM8k	MATH	G-Hard	SVAMP	TabMWP	ASDiv	MAWPS	Avg.
ChatGPT (PAL)	78.6	38.7	67.6	77.8	79.9	81.0	89.4	73.3
GPT-4 (PAL)	97.0	69.7	77.6	94.8	95.9	92.6	97.7	89.3
CodeLLama	34.0	16.6	33.6	59.0	61.4	79.6	-	-
MAmmoTH-7B-Mistral	75.0	40.0	-	-	-	-	-	-
MathCoder-7B-CL	67.8	30.2	-	70.7	-	-	-	-
ToRA-7B-Code	72.6	44.6	56.0	70.4	51.6	78.7	91.3	66.5
MARIO-OVM-7B	74.5	47.7	-	-	-	-	-	-
MMOS-CODE-7B	73.9	44.3	-	76.4	-	78.6	-	-
OpenMath-Mistral-7B	80.2	44.5	63.7	82.4	70.0	82.7	95.4	74.1
Rho-1-Math-7B-Code	81.3	51.8	63.1	80.8	70.1	85.5	94.5	75.3
JiuZhang3.0-7B (Ours)	<u>82.4</u>	<u>53.0</u>	<b>64.9</b>	<b>89.2</b>	<u>75.6</u>	<b>88.3</b>	<u>96.6</u>	78.6
JiuZhang3.0-8B (Ours)	<b>82.9</b>	<b>53.4</b>	<u>64.4</u>	<b>89.2</b>	<b>79.9</b>	<u>87.5</u>	<b>97.3</b>	<b>79.2</b>

our used synthesis model is a much smaller 7B LLM, which has been trained by distilling the data synthesis capability from GPT-4. Thus, it can guarantee the quality of the synthetic data, and helps JiuZhang3.0 models perform the best across most of the dataset. The higher quality also reduces the data amount requirement for pre-training. Owing to our designed high-value data selection strategy, we can also reduce the times of invoking GPT-4 API for knowledge distillation. As noted in Figure 1, the total cost of our approach is nearly only 20% of the compared baselines, indicating its efficiency.

The results of other datasets with different data formats or related to other fields are shown in Table 2. As the listed datasets focus on evaluating the different aspects, the performance of LLMs also differ a lot. For TabMWP, AQuA, and OCW-Math, our JiuZhang3.0-8B and JiuZhang3.0-8x7B achieve the best performance. The three datasets require the understanding of table data, algebra, and undergraduate-level science knowledge respectively, which may have been covered in our synthetic math problems guided by the multi-source corpus. However, our JiuZhang3.0 models perform not well on MMLU-STEM. It indicates the shortcoming of our approach that our prompts and math-related texts might not well cover the knowledge from other subjects.



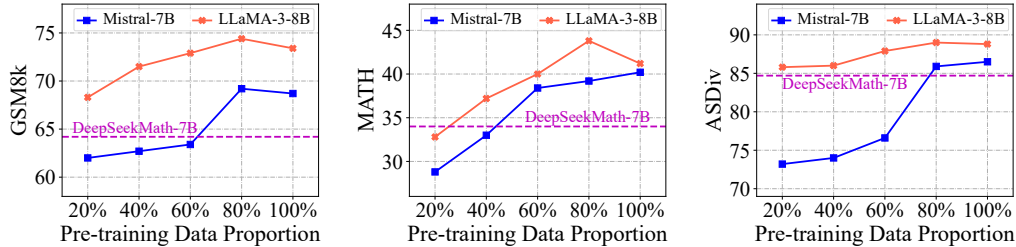


Figure 3: Performance changes with the increasing of the pre-training data proportion for our approach. We also show the best-performed base LLM DeepSeekMath-7B using dashed line.

Table 4: Ablation and variation studies in the natural language reasoning setting. We randomly sample 100k synthetic data and 50k instruction data for efficient test.

Variation	Models	GSM8k	MATH	ASDiv	CARP
-	Ours	78.6	<b>32.8</b>	<b>84.5</b>	36.2
Ablation	w/o Prompt Set	76.9	27.8	81.4	34.5
	w/o Math-related Texts	76.4	28.6	83.8	31.9
	w/o Boosting Retraining	77.9	31.6	83.8	34.7
	w/o Value Estimation	78.8	31.0	83.4	34.3
	w/o using GPT-4 for Boosting	79.0	28.4	82.9	<b>36.3</b>
Synthesis LLM	- ChatGPT	77.0	26.6	83.0	34.3
	- Mixtral-8×7B	77.6	26.8	82.9	33.1
	- DeepSeekMath-RL-7B	77.1	27.2	82.5	32.9
	- LLaMA-3-8B-Instruct	75.7	26.2	81.5	31.4
Data Selection	- Random Sampling	78.8	31.0	83.4	34.3
	- Perplexity	77.5	30.8	83.1	<b>36.3</b>
	- Reward Model	78.0	31.6	84.3	34.8
	- One-shot ICL	<b>79.2</b>	30.2	83.3	36.0

**Tool Manipulation.** The results are shown in Table 3. We can see that our JiuZhang3.0-7B and 8B models outperform all the baseline methods by a large margin, indicating the effectiveness of our approach in this setting. The reason is that we synthesize massive math problems in this format, which can teach JiuZhang3.0 models to accurately utilize tools by generating programs. Besides, the mixed synthetic math problem from the natural language reasoning setting can also benefit the required capabilities for this setting. Different from the natural language reasoning setting, JiuZhang3.0-8B (based on LLaMA-3-8B) performs better than the 7B version (based on Mistral-7B). The reason may be that LLaMA-3-8B owns stronger code synthesis and tool manipulation capability than Mistral-7B.

### 4.3 Further Analysis

**Performance w.r.t. Pre-training Data Amount.** In this part, we study how the scaling of synthetic data amount affects the model performance. We train Mistral-7B and LLaMA-3-8B using varying ratios of our synthetic entire dataset, *i.e.*, 20%, 40%, 60%, 80%, 100%, and report the performance on GSM8k, MATH, and ASDiv under the natural language reasoning setting. For comparison, we also show the results of the best-performed base LLM, *i.e.*, DeepSeekMath-Base-7B.

As shown in Figure 3, with the increasing of the training data ratio, the performance of our model improves consistently. Based on Mistral-7B, it can outperform the best-performed baseline using only 80% or 60% of the pre-training data, indicating the high quality of our synthetic pre-training data. Based on LLaMA-3-8B, it can perform better than the baseline using 40% or even 20% data, and the performance is consistently better than using Mistral-7B. It demonstrates that LLaMA-3-8B can better adapt into our synthetic data. Besides, the performance of our model can surpass the baseline more on MATH, which is a very complex dataset consisting of competitive problems, exhibiting the superiority of our method for improving the advanced mathematical reasoning capability.

**Ablation Study.** We conduct the ablation study to verify the effectiveness of key components in our proposed method. We test the following variations based on our approach, *i.e.*, (1) *w/o Prompt Set*: uses a simple prompt for guiding data synthesis instead of our crafted prompt set; (2) *w/o Math-related Texts*: directly synthesizes the math problems without math related texts; (3) *w/o Boosting Retraining*: uses the data synthesis model without retraining; (4) *w/o Value Estimation*: ignores the estimated value but randomly samples the instances for boosting training; (5) *w/o using GPT-4 for Boosting*: directly uses the high-value instance for boosting data synthesis model instead of using GPT-4. Limited by the computing resource, we conduct the ablation study under the natural language reasoning setting, and use 100k synthetic instances and randomly select 50k instructions from the instruction set. We report the results on GSM8K, MATH, ASDiv and CARP-en.

As shown in Table 4, all the variations mostly underperform the original model, indicating the effectiveness of all the components. Besides, the variation *w/o using GPT-4 for Boosting* performs slightly better in GSM8k and CARP, but degrades a lot in MATH (32.8 $\rightarrow$ 27.8). A possible reason is that it can benefit from the selected high-value data. But without the help of GPT-4, it can not synthesize helpful complex math problems for the competitive problems within MATH dataset.

**Variation Study for Data Synthesis LLMs.** To verify the effectiveness of our trained data synthesis LLM, we conduct the variation study using other existing LLMs for synthesizing the pre-training data. We select the following four LLMs, *i.e.*, ChatGPT, Mixtral-8 $\times$ 7B, DeepSeekMath-RL-7B, and LLaMA-3-8B-Instruct to replace our data synthesis LLM. We follow the efficient test setting in the ablation study, and report the results on GSM8K, MATH, ASDiv and CARP-en.

As shown in Table 4, all the variations mostly perform worse than the original model. It demonstrates that existing LLMs without adapted training might not be suitable to directly synthesize the data for pre-training. Besides, the performance of all the variation degrades a lot in MATH, which consists of complex competitive problems. It indicates that these existing LLMs are hard to synthesize the data that is useful for improving the performance in solving complex math problems.

**Variation Study for Data Selection Strategies.** To study the effectiveness of our gradient-based data selection strategy, we implement the following variations that replace it by other methods, *i.e.*, (1) *Random Sampling*: randomly samples the same number of instances; (2) *Perplexity*: selects the instances with lowest perplexity evaluated by Mistral-7B; (3) *Reward Model*: uses a well-trained reward model [36] for scoring; (4) *One-shot ICL*: concatenates the synthetic math problem and solution with the downstream task data to construct the one-shot in-context learning (ICL) example, and computes the decrease of loss as the estimated value [62]. We follow the efficient test setting.

As shown in Table 4, our original model mostly performs the best among all the variations, indicating the superiority of our gradient-based strategy. Whereas, the variation using one-hot ICL performs relatively better than others, and achieves the best performance on GSM8k. As the problems in GSM8k typically require more natural language reasoning steps, the ICL loss can well detect the instances with helpful context for solving these problems. However, it performs not well on MATH, where the math problems are complex and require using more math symbols and formulas.

## 5 Conclusion

In this paper, we proposed an efficient way to improve the mathematical reasoning of LLMs, where we trained a small LLM to synthesize sufficient high-quality math problems for pre-training. Concretely, we crafted a set of prompts that cover the knowledge and difficulty levels of human education stages, and selected the high-value math-related texts for downstream math-related tasks via the gradient-based strategy. Then, we fed them into GPT-4 to create the knowledge distillation dataset, which can better teach the data synthesis model to generate diverse and useful math problems. We utilized the synthetic data to pre-train JiuZhang3.0, and the whole process only required to invoke GPT-4 API 9.3k times and pre-train on 4.6B data. JiuZhang3.0 achieved state-of-the-art performance on several datasets under the natural language reasoning and tool manipulation settings, surpassing competitive LLMs that requires much larger cost on data synthesis or pre-training.

## Acknowledgments and Disclosure of Funding

This work was partially supported by Beijing Natural Science Foundation under Grant No. L233008 and 4222027, National Natural Science Foundation of China under Grant No. 62222215.

## References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.
- [2] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.
- [3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. A survey of large language models. *ArXiv*, abs/2303.18223, 2023.
- [4] Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning for mathematical reasoning. *ArXiv*, abs/2212.10535, 2022.
- [5] Zhihong Shao, Peiyi Wang, Qihao Zhu, R. X. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.
- [6] Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *ArXiv*, abs/2403.02333, 2024.
- [7] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *ArXiv*, abs/2310.10631, 2023.
- [8] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858, 2022.
- [9] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning. *ArXiv*, abs/2402.06332, 2024.
- [10] Long Long Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zheng Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *ArXiv*, abs/2309.12284, 2023.
- [11] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *ArXiv*, abs/2402.14830, 2024.
- [12] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. *ArXiv*, abs/2309.17452, 2023.
- [13] Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhao Chen. Mammoth2: Scaling instructions from the web. 2024.
- [14] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206, 2023.

- [15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.
- [16] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, March 2024.
- [17] Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024.
- [18] Meta. Introducing meta llama 3: The most capable openly available llm to date, April 2024. <https://ai.meta.com/blog/meta-llama-3/>.
- [19] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024.
- [20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.
- [21] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601, 2023.
- [22] K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Aguerre y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomavsev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172 – 180, 2022.
- [23] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv*, abs/2211.09085, 2022.
- [24] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.
- [25] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.
- [26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.
- [27] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.
- [28] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

- [29] Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.
- [30] Wenhua Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv*, abs/2211.12588, 2022.
- [31] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.
- [32] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022.
- [33] Beichen Zhang, Kun Zhou, Xilin Wei, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji rong Wen. Evaluating and improving tool-augmented computation-intensive math reasoning. *ArXiv*, abs/2306.02408, 2023.
- [34] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [35] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. 2023.
- [36] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing llm reasoning generalists with preference trees. *ArXiv*, abs/2404.02078, 2024.
- [37] Wayne Xin Zhao, Kun Zhou, Zheng Gong, Beichen Zhang, Yuanhang Zhou, Jing Sha, Zhigang Chen, Shijin Wang, Cong Liu, and Ji rong Wen. Jiuzhang: A chinese pre-trained language model for mathematical problem understanding. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.
- [38] Wayne Xin Zhao, Kun Zhou, Beichen Zhang, Zheng Gong, Zhipeng Chen, Yuanhang Zhou, Ji rong Wen, Jing Sha, Shijin Wang, Cong Liu, and Guoping Hu. Jiuzhang 2.0: A unified chinese pre-trained language model for multi-task mathematical problem solving. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.
- [39] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *ArXiv*, abs/2309.05653, 2023.
- [40] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *ArXiv*, abs/2403.04706, 2024.
- [41] Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew Chi-Chih Yao. Augmenting math word problems via iterative question composing. *ArXiv*, abs/2401.09003, 2024.
- [42] Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, 2024.
- [43] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need. *ArXiv*, abs/2312.02120, 2023.
- [44] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data for language models. *ArXiv*, abs/2404.07503, 2024.
- [45] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *ArXiv*, abs/2304.12244, 2023.

- [46] Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. Reformatted alignment. *ArXiv*, abs/2402.12219, 2024.
- [47] Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. Better synthetic data by retrieving and transforming existing datasets. 2024.
- [48] Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. Let’s synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. *ArXiv*, abs/2310.13671, 2023.
- [49] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *ArXiv*, abs/2308.09583, 2023.
- [50] Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *ArXiv*, abs/2402.16352, 2024.
- [51] Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *ArXiv*, abs/2402.13064, 2024.
- [52] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv*, abs/2305.11738, 2023.
- [53] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *ArXiv*, abs/2402.04333, 2024.
- [54] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.
- [55] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [56] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Mądry. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27074–27113, 2023.
- [57] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984.
- [58] Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. *ArXiv*, abs/2310.06786, 2023.
- [59] Zengzhi Wang, Rui Xia, and Pengfei Liu. Generative ai for math: Part i - mathpile: A billion-token-scale pretraining corpus for math. *ArXiv*, abs/2312.17120, 2023.
- [60] Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew Chi-Chih Yao. Autonomous data selection with language models for mathematical texts. 2024.
- [61] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.
- [62] Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiayi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One shot learning as instruction data prospector for large language models. *ArXiv*, abs/2312.10302, 2023.

- [63] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021.
- [64] Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *ArXiv*, abs/2312.08935, 2023.
- [65] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *ArXiv*, abs/2402.10176, 2024.
- [66] Zui Chen, Yezeng Chen, Jiaqi Han, Zhijie Huang, Ji Qi, and Yi Zhou. An empirical study of data ability boundary in llms’ math reasoning. *ArXiv*, abs/2403.00799, 2024.
- [67] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. *ArXiv*, abs/2106.15772, 2020.
- [68] Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. Mawps: A math word problem repository. In *North American Chapter of the Association for Computational Linguistics*, 2016.
- [69] Arkil Patel, S. Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *North American Chapter of the Association for Computational Linguistics*, 2021.
- [70] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, 2017.
- [71] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv*, abs/2304.06364, 2023.
- [72] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020.
- [73] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858, 2022.
- [74] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *ArXiv*, abs/2304.09842, 2023.
- [75] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *ArXiv*, abs/2309.16609, 2023.
- [76] DeepSeek-AI Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wen-Hui Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren,

- Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Jun-Mei Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Min Tang, Bing-Li Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Yu Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yi Xiong, Hanwei Xu, Ronald X Xu, Yanhong Xu, Dejian Yang, Yu mei You, Shuiping Yu, Xin yuan Yu, Bo Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghu Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism. *ArXiv*, abs/2401.02954, 2024.
- [77] Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *ArXiv*, abs/2404.02893, 2024.
- [78] Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295, 2024.
- [79] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, J'eremy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- [80] Zheng-Wen Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. *ArXiv*, abs/2404.07965, 2024.
- [81] Ethan Chern, Haoyang Zou, Xuefeng Li, Jiewen Hu, Kehua Feng, Junlong Li, and Pengfei Liu. Generative ai for math: Abel. <https://github.com/GAIR-NLP/abel>, 2023.
- [82] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *ArXiv*, abs/2310.03731, 2023.
- [83] Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. Mario: Math reasoning with code interpreter output - a reproducible pipeline. *ArXiv*, abs/2401.08190, 2024.
- [84] Xinzhe Ni, Yeyun Gong, Zhibin Gou, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Exploring the mystery of influential data for mathematical reasoning. *ArXiv*, abs/2404.01067, 2024.
- [85] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chaochao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *ArXiv*, abs/2404.06395, 2024.



Table 5: The price of the potential service during the training procedure.

Items	Details
OpenAI API	30.00 USD per 1M tokens for input, and 60.00 USD per 1M tokens for output.
AWS GPU Server	40.96 USD per $8 \times$ A100 per hour.

Table 6: The estimated cost of different LLMs using the official GPT-4 API and 8 nodes of  $8 \times$  A100 GPU servers for training.

Models	#API Calling (Token)			#Server Time (Hour)		Total Expenses
	Input	Output	Data Num.	Synthesizing	Training	
KPMath-DSMath-7B	1090	218	865K	-	-	39,599 USD
DeepSeekMath-7B-RL	-	-	-	-	160	52,428 USD
JiuZhang3.0-7B	300	877	10K	14	10	8,480 USD

## A Cost Estimation

To estimate the expenses of previous work and our proposed JiuZhang3.0, we survey the price of potential service during the entire procedure, including calling the OpenAI API for GPT-4 and renting the AWS GPU server for LLMs training. The details of the price are presented in Table 5. For a fair comparison, we assume that GPT-4 is utilized to synthesize training data, and 8 nodes of  $8 \times$  A100 GPU servers (64 GPUs in total) are leveraged for LLMs training.

For our data synthesis process, the average length of the prompting (including selected math-related texts) is about 300 tokens, and the average length of problems and solutions is 877 tokens. Under the setting of 64 GPUs, we spend 4 hours selecting valuable data for natural language reasoning and tool manipulation in total, and 10 hours synthesizing the 4.6B pre-training corpus. We train the JiuZhang3.0-7B model for 10 hours, including both pre-training and fine-tuning.

As the dataset and the construction process of KPMath are not publicly available, we adopt the average length of synthesis prompts, problems, and solutions in MetaMathQA for estimating the input and the output tokens, respectively. We estimate the training time on 120B tokens for a 7B model as 160 hours. Since the training details in the RLHF stage of DeepSeekMath-7B-RL are not publicly available, we do not count the fine-tuning cost for the KPMath model and the DeepSeekMath model. In this case, we estimate the expenses of the LLMs training process as follows,

$$\begin{aligned} \text{API Expenses} = & (\text{Avg. Input Length} \times \text{API Input Price} \\ & + \text{Avg. Output Length} \times \text{API Output Price}) \times \text{Num Data}, \end{aligned} \tag{6}$$

$$\text{Server Expenses} = \text{Num Nodes} \times \text{Price per Node} \times \text{Training Time}, \tag{7}$$

$$\text{Total Expenses} = \text{API Expenses} + \text{Server Expenses}. \tag{8}$$

The details and estimation of the expenses for different LLMs are present in Table 6.

## B Fine-tuning Data

After pre-training, we collect a set of open-source math-related instructions to fine-tune JiuZhang3.0. For natural language reasoning, we collect the training sets of MATH [63], GSM8k [34], CARP [33], and open-source synthetic datasets based on them, *i.e.*, MetaMATH [10], MMIQC [41], Math-Shepherd (without problems from original MATH test sets) [64], Orca-MATH [11]. Besides, we also collect the positive examples from the PRM800k dataset (without problems from original MATH test sets) [35], and the TAL-SCQ5K dataset consisting of multi-choice questions<sup>3</sup>. Whereas, we observe that the varying data styles in the above datasets might cause the LLM outputs to be irregular. Thus, we utilize a unified prompt DeepSeekMath-7B-RL, to synthesize 700k solutions for the problems from the above datasets.

<sup>3</sup><https://github.com/math-eval/TAL-SCQ5K>

For tool manipulation, we use the synthetic datasets, *i.e.*, OpenMathInstruct-1 [65] and MMOS [66], consisting of a mixture of text reasoning and code blocks executed by a Python interpreter.

## C Evaluation Datasets

We test our JiuZhang3.0 and baseline methods in the following two settings for evaluating the mathematical reasoning capability.

- *Natural Language Reasoning*: we prompt LLMs to perform multi-step reasoning via natural language, and select the following publicly available datasets: GSM8k [34] contains grade school math problems to test the basic arithmetic and reasoning ability. MATH [63] and CARP-en [33] consist of complex competition-level problems, and CARP-en is the English version of the original CARP dataset. ASDiv [67], MAWPS [68] and SVAMP [69] are grade-school math word problem (MWP) datasets, and SVAMP focuses on the robust reasoning ability. Besides, we also consider the following datasets with different data formats or related to other interdisciplinary field, *i.e.*, TabMWP, AQuA, SAT-Math, MathQA, MMLU-STEM. AQuA [70], SAT-Math [71], MMLU-STEM [72] are composed of multiple-choice questions for human exams across math and other STEM disciplines. OCW-Math [73] is a challenging dataset containing undergraduate-level math and science problems.
- *Tool Manipulation*: we prompt LLMs to manipulate external tools via Python to solve the problems, and select GSM8k, MATH, GSM-Hard, SVAMP, TabMWP, ASDiv, and MAWPS for testing. GSM-Hard [27] replaces the numbers in the questions of GSM8K with larger numbers to increase the difficulty of calculation. TabMWP [74] is an open-domain MWP dataset containing tabular data.

## D Baseline Methods

We consider diverse types of baseline methods for comparison.

- *Closed-source LLMs*: ChatGPT and GPT-4 [2];
- *Larger LLMs (>20B)*: Qwen-1.5-110B [75], Qwen-1.5-72B [75], Deepseek-LM-67B [76], Mixtral-8×7B [19], Llemma-34B [7], Intern-Math-20B [9], MAMmoTH2-8×7B-Plus [13], ChatGLM-Math-32B [77];
- *Smaller LLMs (<10B)*: DeepSeek-7B [76], Qwen-1.5-7B [75], Mistral-7B [61], LLaMA-3-8B [18], Gemma-7B [78], and CodeLLama [79];
- *LLMs pre-trained on Math Corpus (<10B)*: Llemma-7B [7], InternLM-Math-7B [9], Rho-1-Math-7B [80], DeepSeekMath-7B [5];
- *LLMs fine-tuned on Math Instructions (<10B)*: MetaMath-Mistral-7B [10], WizardMath-7B-1.1 [49], Abel-7B-002 [81], Mistral-7B-MMIQC [41], Math-Shepherd-Mistral-7B-RL [64], DeepSeekMath-7B-Instruct [5], DeepSeekMath-7B-RL [5], Llama-3-8B-Instruct [18], MAMmoTH2 [13], KPMath-DSMath-7B [6];
- *LLMs fine-tuned on tool-augmented math instructions (<10B)*: MAMmoTH-7B-Mistral [39], MathCoder-7B-CL [82], ToRA-7B-Code [12], MARIO-OVM-7B [83], MMOS-CODE-7B [66], OpenMath-Mistral-7B [65], Rho-1-Math-7B-Code [80].

Our evaluation framework and in-context examples follow the existing work [6, 12, 84]. For general and math domain base models, we adopt the few-shot prompting method. For fine-tuned models, we adopt the zero-shot prompting method for open-ended natural language reasoning and tool manipulation tasks, and the few-shot prompting method for multiple choice problems. We cite the performance results reported in existing work [6, 12, 39, 65, 66, 77, 80, 82, 83].

## E Implementation Details

**Data Synthesis Models.** We train two data synthesis models for the natural language reasoning and tool manipulation settings, respectively. We first initialize them by training DeepSeekMath-7B-RL with 4k and 1.3k KD datasets, respectively. Then, we utilize the synthesis models to generate 100k problem-solution pairs for each one. During value estimation, we adopt the training set of GSM8k and MATH for natural language reasoning, and the 5k subset from a mixture of OpenMathInstruct

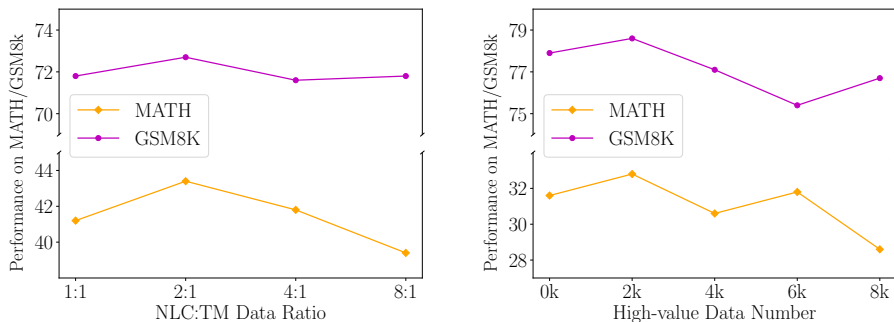


Figure 4: Hyper-parameter tuning results of the pre-training data proportion and high-value data amount, under the tool manipulation and natural language reasoning settings, respectively.

and MMOS for tool manipulation, as the instances from downstream math-related tasks. We select 2k the most valuable math texts in each setting for GPT-4 to boost the quality, and then add them into the KD dataset. During training, following existing work [14], we adopt a cosine learning rate schedule with a 0% warm-up ratio and select a learning rate of  $1e-5$  for 5 epochs and 10 epochs for natural language reasoning and tool manipulation, respectively.

**JiuZhang3.0.** Before training, we first filter out the instance from the synthetic dataset that have 10-grams overlap to the test set data, and also deduplicate data and the synthetic tool manipulation data containing unexecutable code. Then, we follow existing work that trains 7B, 8B, and  $8\times 7B$  versions based on Mistral-7B, LLaMA-3-8B, and Mixtral- $8\times 7B$  [13]. During training, we first pre-train it on our synthetic 4.6B math problem-solution pairs and then fine-tune it on the multi-source instruction set. We reuse the optimizer to initialize the fine-tuning stage and adopt the Warmup-Stable-Decay learning rate scheduler [85] with 3% warm-up ratio and 85% stable training ratio for 1 epoch in the whole training process. We set the maximum learning rate to  $1e-5$  and the minimum learning rate to  $1e-6$  with a total batch size of 512. To boost the training efficiency, we pack multiple instances in the same context window of the model and modify the attention to avoid mutual interference among difference instances. The maximum length of model is set to 2048. We train all models with BFloat16 numerical format, Flash Attention 2.0, DeepSpeed Stage 2 for 7B and 8B models, and Stage 3 for  $8\times 7B$  models.

## F Hyper-parameter Tuning

In this part, we conduct the experiments about tuning two important hyper-parameters, *i.e.*, pre-training data proportion, and the number of high-value data.

**Pre-train Data Proportion.** In the synthetic pre-training data proportion, we should determine the ratio between the data from the natural language reasoning and tool manipulation settings (NLC : TM). We set it to 1:1, 2:1, 4:1, and 8:1, and control the total data amount unchanged (200k) for comparison. We follow the efficient test setting in Section 4.3, and report the MATH and GSM8k results under the tool manipulation setting, as there is only slight performance changes on natural language reasoning setting. As shown in Figure 4, 2:1 is more suited and leads to better performance than other proportions. The reason may be that smaller or larger proportion for tool manipulation data would cause underfitting or overfitting on the tool manipulation data, affecting the corresponding capability.

**Number of Selected High-Value Data.** We also study the effect of changing the number of selected high-value data according to the ranking, as more data requires larger cost for invoking the GPT-4 API. We set it to 2k, 4k, 6k, and 8k, and do not change other settings for fair comparison. We follow the efficient test setting in Section 4.3. As shown in Figure 4, using the top 2k high-value data achieves the best performance, and more data even causes the performance degradation. It indicates

that the data quality is more important than data amount for boosting the data synthesis capability, as using more data would inevitably involve low-quality data.

## **G Limitation**

First, although we train a strong yet small LLM that can synthesize high-quality math problems for training, its capabilities on synthesizing the data for other domains or tasks might be relatively weaker. The reason is that we only use the math-related data to train it. In future work, we will try to train a general-purpose model, to enable the data synthesis for other requirements. Second, in this work, we only focus on mathematical reasoning capability, and our trained JiuZhang3.0 is also mainly for solving math problems. Limited by the computation resource, we do not test its performance on other complex reasoning tasks, *e.g.*, planning, commonsense reasoning. We will also conduct the corresponding experiments in the future. Third, the 4.6B pre-training data is still a large scale for training, and we do not perform data filter to control its quality. Future work should focus on reducing its scale by proposing better data filter strategy. Forth, we only utilize GPT-4 for knowledge distillation, but do not use other well-performed LLMs, *e.g.*, Claude 3, GLM-4, and the latest GPT-4o. More experiments should be conducted on these LLMs to study the effect of the teacher LLMs. Fifth, for cost estimation, our reported results are all estimated according to our experience, without the re-implementation of these methods. We also appeal to report the true cost of training LLMs in existing work.

## H Case Study

### Good Case 1

**Type:** Natural Language Reasoning – Grade School

**Math-Related Texts:**

**Question**

The Huka Falls on the Waikato River is one of New Zealand's most visited natural tourist attractions (see Figure 12.30). On average the river has a flow rate of about 300,000 L/s. At the gorge, the river narrows to 20 m wide and averages 20 m deep. (a) What is the average speed of the river in the gorge? (b) What is the average speed of the water in the river downstream of the falls when it widens to 60 m and its depth increases to an average of 40 m?

**Question Image**

1. 0.75 m/s

2. 0.13 m/s

**Solution Video#** OpenStax College Physics Solution, Chapter 12, Problem 5 (Problems & Exercises) (1:14)

**View sample solution##** Calculator Screenshots

**Video Transcript**

This is College Physics Answers with Shaun Dychko. At the gorge where the Waikato River of New Zealand narrows to a width of 20 meters and a height of 20 meters, we're going to figure out what its speed is. And so the volume flow rate is equal to the cross sectional area of the river multiply by its

**Synthesis QA:**

**Question:** The Waikato River flows through a gorge where it narrows down to a width of 20 meters and a depth of 20 meters. The average flow rate of the river is 300,000 liters per second.

**\*\*Question:\*\*** What is the average speed of the river in this gorge?

**Answer:**

To find the average speed of the river in the gorge, we need to use the formula for the volume flow rate:

$$\text{Volume Flow Rate} = \text{Cross-Sectional Area} \times \text{Speed}$$

First, let's calculate the cross-sectional area of the river:

$$\text{Cross-Sectional Area} = \text{Width} \times \text{Depth}$$

Given:

- Width = 20 meters

- Depth = 20 meters

So,

$$\text{Cross-Sectional Area} = 20 \text{ meters} \times 20 \text{ meters} = 400 \text{ square meters}$$

Now, we know the volume flow rate and the cross-sectional area. The volume flow rate ( $Q$ ) is 300,000 liters per second, which we need to convert into cubic meters because the area is in square meters:

$$1 \text{ liter} = 1 \text{ cubic decimeter} = 0.001 \text{ cubic meters}$$

So,

$$Q = 300,000 \text{ liters per second} = 300,000 \times 0.001 \text{ cubic meters per second} = 300 \text{ cubic meters per second}$$

We can now use the formula for the volume flow rate to find the speed ( $v$ ):

$$Q = \text{Cross-Sectional Area} \times v$$

Rearranging to solve for  $v$ :

$$v = \frac{Q}{\text{Cross-Sectional Area}}$$

Substituting the values we have:

$$v = \frac{300 \text{ cubic meters per second}}{400 \text{ square meters}} = 0.75 \text{ meters per second}$$

So, the average speed of the river in the gorge is **0.75 meters per second**

## Good Case 2

**Type:** Natural Language Reasoning – Middle

**Math-Related Texts:**

Successive parabolic interpolation

Successive parabolic interpolation is a technique for finding the extremum (minimum or maximum) of a continuous unimodal function by successively fitting parabolas (polynomials of degree two) to a function of one variable at three unique points or, in general, a function of  $n$  variables at  $1+n(n+3)/2$  points, and at each iteration replacing the oldest point with the extremum of the fitted parabola.

Advantages

Only function values are used, and when this method converges to an extremum, it does so with an order of convergence of approximately 1.325. The superlinear rate of convergence is superior to that of other methods with only linear convergence (such as line search). Moreover, not requiring the computation or approximation of function derivatives makes successive parabolic interpolation a popular alternative to other methods that do require them (such as gradient descent and Newton's method).

Disadvantages

On the other hand, convergence (even

**Synthesis QA:**

**Question:** Alex and Jamie are working on a science project and need to find the maximum height of a rocket launched upward.

They record the height of the rocket at three different times:

- At  $t = 0$  seconds, the height is 0 meters.
- At  $t = 5$  seconds, the height is 25 meters.
- At  $t = 10$  seconds, the height is 40 meters.

To find the maximum height, Alex and Jamie decide to use successive parabolic interpolation. They start by fitting a parabola to these three points and then finding the maximum point of this parabola.

1. Write the general form of a quadratic (parabolic) equation:  $y = ax^2 + bx + c$ . 2. Use the three points to create a system of equations to find the coefficients  $a$ ,  $b$ , and  $c$ . 3. Find the vertex of the parabola to determine the maximum height.

**Answer:**

1. **General Form of a Quadratic Equation:**

The general form of a quadratic equation is:

$$y = ax^2 + bx + c$$

2. **Creating a System of Equations:**

We have three points:

$$(0, 0), (5, 25), (10, 40)$$

Plugging these points into the equation  $y = ax^2 + bx + c$ :

- For  $(0, 0)$ :

$$0 = a(0)^2 + b(0) + c \implies c = 0$$

- For  $(5, 25)$ :

$$25 = a(5)^2 + b(5) + 0 \implies 25 = 25a + 5b \quad (\text{Equation 1})$$

- For  $(10, 40)$ :

$$40 = a(10)^2 + b(10) + 0 \implies 40 = 100a + 10b \quad (\text{Equation 2})$$

3. **Solving the System of Equations:**

Use Equation 1:

$$25 = 25a + 5b \quad (\text{Equation 1})$$

Simplify:

$$5 = 5a + b \quad (\text{Equation 3})$$

Use Equation 2:

$$40 = 100a + 10b \quad (\text{Equation 2})$$

Simplify:

$$4 = 10a + b \quad (\text{Equation 4})$$

Subtract Equation 4 from Equation 3:

$$(5a + b) - (10a + b) = 5 - 4 \implies -5a = 1 \implies a = -\frac{1}{5}$$

Plug  $a = -\frac{1}{5}$  back into Equation 3:

$$5 = 5\left(-\frac{1}{5}\right) + b \implies 5 = -1 + b \implies b = 6$$

## Good Case 2

4. **\*\*Vertex of the Parabola:\*\***

The vertex of a parabola  $y = ax^2 + bx + c$  is given by:

$$x = -\frac{b}{2a}$$

Here,  $a = -\frac{1}{5}$ ,  $b = 6$ , and  $c = 0$ :

$$x = -\frac{6}{2\left(-\frac{1}{5}\right)} = -\frac{6}{-\frac{2}{5}} = 15$$

So, the maximum height occurs at  $t = 15$  seconds. However, this  $t$  value exceeds the range of our data points. In practical interpolation methods, this would indicate a correction point is needed beyond the given data. But for demonstration, let's assume  $t = 15$  is within bounds (extrapolation for students):

Plug  $t = 15$  back into the original equation with adjusted coefficients:

$$y = -\frac{1}{5}(15)^2 + 6(15) = -\frac{1}{5}(225) + 90 = -45 + 90 = 45$$

Thus, the maximum height is **\*\*45 meters\*\***

## Good Case 3

**Type:** Natural Language Reasoning – High

**Math-Related Texts:**

Let  $k$  be a positive integer. Find all polynomials with real coefficients which satisfy the equation  $P(P(x)) = (P(x))^k$ . Let  $k$  be a positive integer. Find all polynomials with real coefficients which satisfy the equation

$$P(P(x)) = (P(x))^k.$$

I simply don't even know how to think about this problem. I've tried simple stuff just to get my head on the problem. For example for  $P(x) = x^n$  I have  $P(P(x)) = (P(x))^n$ , and I think that any polynomial  $P(x) = x^n + x^{n-1} + \dots + c$  can't be a solution as I would have  $P(x) = P(x)q_1(x) + R$ . After that I simply stare at the problem. Can you guys give some help? Note: I would like to understand how to tackle these kind of problems, so I would be really grateful if you would explain the thinking process behind the solution. (This is optional, so feel free to give an answer as you prefer.) Thanks in advance.

First we note that only the constant polynomials  $P \equiv 0$  and  $P \equiv 1$ , as well as  $P \equiv -1$  for odd  $k$ , satisfy the equation. (This holds true for  $k > 1$ ; for  $k = 1$  any constant polynomial will do.) Now if  $P$  is not constant, then the range  $Y = \{P(x) : x \in \mathbb{R}\}$  of  $P$  is an infinite set. For every  $y \in Y$  we have  $P(y) = y^k$ , which means that the polynomial  $Q(x) = P(x) - x^k$  is zero on  $Y$ . Since  $Y$  is infinite, this implies that  $Q$  is the zero polynomial. In conclusion,  $P(x) = x^k$ .

**Synthesis QA:**

**Question:** Consider the polynomial  $P(x) = x^2$ . We want to find out how  $P$  behaves when we apply it twice,  $P(P(x))$ , and compare it with  $P(x)$  raised to some power,  $(P(x))^k$ .

Let's explore the following scenario:

Given  $P(x) = x^2$ , find all values of  $k$  such that the equation  $P(P(x)) = (P(x))^k$  holds for all real numbers  $x$ . **Answer:** First, we evaluate  $P(P(x))$ :

$$P(P(x)) = P(x^2) = (x^2)^2 = x^4$$

Next, we need to express  $P(x)$  raised to some power  $k$ :

$$(P(x))^k = (x^2)^k = x^{2k}$$

We want  $P(P(x)) = (P(x))^k$ :

$$x^4 = x^{2k}$$

This equation holds for all real numbers  $x$  if and only if the exponents are equal. Therefore, we set up the following equation:

$$4 = 2k$$

Solving for  $k$ :

$$k = \frac{4}{2} = 2$$

Thus, the equation  $P(P(x)) = (P(x))^k$  holds for  $k = 2$ .

Now let's verify this:

For  $k = 2$ :

$$\begin{aligned} P(P(x)) &= x^4 \\ (P(x))^k &= (x^2)^2 = x^4 \end{aligned}$$

Since both sides of the equation are equal:

$$x^4 = x^4$$

This equation holds true for all real numbers  $x$ .

Therefore, the only value of  $k$  that makes  $P(P(x)) = (P(x))^k$  true for all real numbers  $x$  is  $k = 2$ .

## Good Case 4

**Type:** Tool Manipulation – Grade School

**Math-Related Texts:**

Refer to a unit in the Units package by indexing the name or symbol with the context, for example, bit[information] or bps[information]; or, if the context is indicated as the default, by using only the unit name or symbol, for example, bit or bps. The units of information are defined as follows. A crumb is defined as 2 bits. A nibble is defined as 4 bits.

**Synthesis QA:**

**Question:** In a digital communication system, information is transmitted in units called bits. A crumb is defined as 2 bits, and a nibble is defined as 4 bits. If a user sends a sequence of bits, and the sequence can be broken down into whole crumbs and nibbles, how many different ways can the user send exactly 18 bits?

**Answer:**

```
def count_ways_to_send_bits(total_bits):
    ways = 0
    for crumb in range(total_bits // 2 + 1):
        for nibble in range(total_bits // 4 + 1):
            if crumb * 2 + nibble * 4 == total_bits:
                ways += 1
    return ways
total_bits = 18
ways_to_send = count_ways_to_send_bits(total_bits)
print(ways_to_send)
```

```
““output
5
““
```

There are  different ways the user can send exactly 18 bits.



## Good Case 5

**Type:** Tool Manipulation – Secondary School Competition

**Math-Related Texts:**

The generalization of this is 0:  $I = -\log(p(x))$  Mean and variance of continuous random variables

The mean is the sum of  $x$  \* density probability function at that point.  $\mu = \int_{-\infty}^{\infty} xp(x)$

The variance is the sum of distance to the mean times the density probability function at that point.  $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)$

Esperance of a probability distribution is the value we could expect to have if we repeat some experiment a lot of times. It's the sum of the value of the events ponderated by the probability of the event to occur.

**Synthesis QA:**

**Question:** Consider a continuous random variable  $X$  with probability density function (pdf) given by:  $p(x) =$

$$\begin{cases} \frac{1}{2}e^{-\frac{x}{2}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

1. Calculate the mean ( $\mu$ ) of the distribution.

2. Calculate the variance ( $\sigma^2$ ) of the distribution.

**Answer:**

```
import sympy as sp

# Define the variable and the pdf
x = sp.Symbol('x', real=True, positive=True)
pdf = (1/2) * sp.exp(-x/2)

# Calculate the mean
mean = sp.integrate(x * pdf, (x, 0, sp.oo))

# Calculate the second moment about the origin
second_moment = sp.integrate(x**2 * pdf, (x, 0, sp.oo))

# Calculate the variance
variance = second_moment - mean**2

mean_value = mean.evalf()
variance_value = variance.evalf()

mean_value, variance_value

""output
(2.0, 4.0)
""
```

The mean ( $\mu$ ) of the distribution is , and the variance ( $\sigma^2$ ) of the distribution is .

## Bad Case 1

**Type:** Natural Language Reasoning – AMC12

**Math-Related Texts:**

Julia Robinson Mathematics Festival

The Julia Robinson Mathematics Festival (JRMF) is an educational organization that sponsors locally organized mathematics festivals and online webinars targeting K–12 students. The events are designed to introduce students to mathematics in a collaborative and non-competitive forum.[1][2][3]

AbbreviationJRMF FounderNancy Blachman Location • c/o SEE, 23564 Calabasas Road, Suite 201, Calabasas, CA 91302 Coordinates37°22'48.57"N 121°54'25.98"W Key people Nancy Blachman, Founder Daniel Kline, Executive Director Website-jrmf.org

**History**

In the 1970s, Saint Mary's College of California produced a mathematics contest that was popular with secondary schools throughout the San Francisco Bay Area.[4] In 2005, Nancy Blachman attended an education forum sponsored by the Mathematical Sciences Research Institute (MSRI) and remembered how the Saint Mary's contest had inspired her as a student.[5] Unfortunately, the contest no longer existed. Seeking to possibl

**Synthesis QA:**

**Question:** At the Julia Robinson Mathematics Festival, students are challenged with a geometry puzzle. The problem is as follows: A square  $ABCD$  with side length 10 is inscribed in a circle. Points  $P$  and  $Q$  lie on the circle such that  $P$  is on the same side of line  $AD$  as  $C$ , and  $Q$  is on the same side of line  $AB$  as  $C$ . If  $AP = 16$  and  $AQ = 12$ , find the length of  $PQ$ .

**Answer:** To solve this problem, we'll start by finding the radius of the circle, using the fact that the square is inscribed in the circle. Then, we'll use properties of circles and the given lengths to determine the length of  $PQ$ .

1. **Finding the radius of the circle:**

Since the square  $ABCD$  is inscribed in the circle, the diameter of the circle is equal to the diagonal of the square. The diagonal of a square with side length  $s$  can be calculated using the Pythagorean theorem:

$$\text{Diagonal} = s\sqrt{2}$$

For the given square with side length 10, the diagonal is:

$$\text{Diagonal} = 10\sqrt{2}$$

Thus, the radius of the circle, which is half the diagonal, is:

$$\text{Radius} = \frac{10\sqrt{2}}{2} = 5\sqrt{2}$$

2. **Using the Power of a Point Theorem:**

The Power of a Point Theorem states that for a point  $P$  and a point  $Q$  in a circle with center  $O$  and radius  $r$ , if  $P$  and  $Q$  are on the same side of a chord  $CD$  passing through  $O$ , then the product of the distances from  $P$  and  $Q$  to  $CD$  is equal to the square of the distance from  $O$  to  $CD$ .

In our case,  $CD$  is the diameter of the circle, and  $O$  is the center of the circle. Point  $P$  is 16 units from  $A$  on side  $AD$ , and  $Q$  is 12 units from  $A$  on side  $AB$ .  $AP = 16$  and  $AQ = 12$ . Since  $P$  and  $Q$  are on the circle and the diameter passes through the center  $O$ , the Power of a Point Theorem gives:

$$AP \cdot AQ = OP^2$$

$$16 \times 12 = OP^2$$

$$OP = \sqrt{192} = 8\sqrt{3}$$

3. **Determining the length of  $PQ$ :**

Finally, we use the radius of the circle and the calculated power of point distance to find  $OP$  to determine  $PQ$ . Since  $O$  is the midpoint of  $PQ$  and  $OP = OQ = 8\sqrt{3}$ :

$$PQ = 2 \times OP = 2 \times 8\sqrt{3} = 16\sqrt{3}$$

Thus, the length of  $PQ$  is  $16\sqrt{3}$  units.

This problem challenges students to use geometric principles (properties of circles and squares) and algebraic reasoning to solve for lengths in a geometric configuration, enhancing analytic and problem-solving skills suitable for the AMC 12 level.

## Bad Case 2

**Type:** Tool Manipulation – Secondary School Competition

**Math-Related Texts:**

ONick D. Jan 5 '17 at 14:48

• I guess it comes from a wifi code. Are you a chinese student? –ORenascence5. Jan 6 '17 at 8:50

**HINT:**

Observe that the function is odd and the interval  $[-5, 5]$  is symmetric about 0. See more in this question:

Definite integral of an odd function is 0 (symmetric interval)

$$x^4 + 2x^2 + 1 = (x^2 + 1)^2$$

so we have an integral of the form

$$\int \frac{x^3}{(x^2 + 1)^2} \sin^2 x dx$$

we can clearly see that

$$f(x) = \frac{x^3}{(x^2 + 1)^2} \sin^2 x \implies f(-x) = -f(x)$$

i.e.

...

Guidelines for each section:

**Synthesis QA:**

**Question:** Consider the function  $f(x) = \frac{x^3}{(x^2+1)^2} \sin^2 x$  over the interval  $[-5, 5]$ . We are interested in evaluating the definite integral of this function with respect to  $x$ . Given the properties of the function  $f(x)$ , which is odd, and the symmetric interval about the origin, determine the value of the integral.

**Answer:**

```
import sympy as sp
```

```
# Define the variable
```

```
x = sp.symbols('x')
```

```
# Define the function
```

```
f = (x**3) / (x**2 + 1)**2 * sp.sin(x)**2
```

```
# Since the function is odd over a symmetric interval, the integral is zero
```

```
integral_result = sp.integrate(f, (x, -5, 5))
```

```
print(integral_result)
```

```
""output
```

```
0
```

```
""
```

```
The answer is .
```

## I Prompts

### I.1 Prompts for Natural Language Reasoning Data Synthesis

#### Prompts for Grade School-Level Problem

**Instruction**

Create an age-appropriate math word problem for grade school students based on the provided math content.

**Math Content**

[Math Text Placeholder]

**Guidelines**

[Problem]: Craft a concise math word problem suitable for grade school, focusing on basic arithmetic operations (addition, subtraction, multiplication, division), number sense, simple shapes, or introductory measurements. Use relatable, real-world scenarios appropriate for the age group. Ensure the problem is purely text-based and solvable without images. [Solution]: Provide a clear, step-by-step solution to the problem using simple language that a grade school student could understand. Explain the reasoning behind each step.

### Prompts for Middle School-Level Problem

**Instruction**

Create an middle school level math problem and solution based on the provided math content excerpt.

**Math Content**

[Math Text Placeholder]

**Guidelines**

[Problem]: Create a self-contained problem for middle school student that directly incorporates a concept from the provided math content. Target a difficulty level appropriate for grades 6-8 (ages 11-14), assuming knowledge of arithmetic, pre-algebra, basic probability/statistics, and geometry. Ensure the problem is fully text-based and solvable without images. Use concepts typically covered by the end of 8th grade.

[Solution]: Provide a detailed, step-by-step solution that demonstrates the mathematical reasoning from problem statement to conclusion, around 250-350 words long. Utilize LaTeX formatting for all mathematical expressions. Explain each step to reinforce the underlying math principles being applied.

### Prompts for High-Level Problem

**Instruction**

Inspired by the provided math content extract, create high school-level math problem that combines concepts from at least two math subjects.

**Math Content**

[Math Text Placeholder]

**Guidelines**

[Problem]: Draft a self-contained math problem for high school students based on the given math content. The problem should require knowledge from one of these subjects: Algebra I and II, Pre-Calculus, Calculus, Geometry, Trigonometry, Statistics and Probability. Ensure the problem is fully text-based and solvable without images. Use concepts typically covered by the end of 11th grade.

[Solution]: Provide a detailed, step-by-step solution that demonstrates the mathematical reasoning from problem statement to conclusion, around 250-350 words long. Utilize LaTeX formatting for all mathematical expressions. Explain each step to reinforce the underlying math principles being applied.

### Prompts for High-Level Problem

**Instruction**

Inspired by the math content, create a college-level math problem.

**Math Content**

[Math Text Placeholder]

**Guidelines**

[Problem]: Draft a self-contained, college-level math problem inspired by the math content. It should be intellectually stimulating and designed for an audience familiar with advanced mathematics, such as Calculus, Linear Algebra, Abstract Algebra, etc. Ensure the problem includes all necessary information for solving it. Aim for a problem statement around 100-150 words. [Solution]: Provide a step-by-step solution to your problem, around 250-350 words long. The solution should clearly explain the reasoning, mathematical principles, and steps used. Call out any key theorems or properties being applied at each step.

### Prompts for AMC 8-Level Problem

**Instruction**

Inspired by the provided math content, craft a math problem suitable for the AMC 8 competition, engaging top grade students with a challenge in areas of arithmetic, algebra, counting, geometry, logic, number theory, or probability. Ensure the problem stimulates problem-solving techniques and heuristics within a text-based and solvable format, avoiding advanced subjects like calculus or physics.

**Math Content**

[Math Text Placeholder]

**Guidelines**

[Problem]: Design a compelling, self-contained math problem for AMC 8 contestants inspired by the math content, incorporating elements of basic math disciplines. The problem should be approachable through logical reasoning and fundamental problem-solving strategies. Ensure the problem is fully text-based, solvable without the aid of images, and has a difficulty level appropriate for the AMC 8 competition.

[Solution]: Provide a detailed, step-by-step solution that is educational and tailored to the AMC 8 audience, around 250-350 words long. Explain the logic and reasoning behind each step thoroughly, using clear and age-appropriate language and terminology to facilitate understanding. Highlight the use of problem-solving techniques and heuristics employed in the solution.

## Prompts for AMC 12-Level Problem

### Instruction

Inspired by the provided math content, develop a math problem suitable for the AMC 12 competition, targeting high-school students with capabilities in advanced areas like algebra, geometry, trigonometry, counting, probability, and number theory. The problem should challenge students to utilize sophisticated problem-solving skills and mathematical reasoning. Focus on concepts from the given content, but adapt them into an original problem. Avoid delving into higher-level college mathematics topics beyond the AMC 12 scope.

### Math Content

[Math Text Placeholder]

### Guidelines

[Problem]: Formulate a compelling and complex math problem for AMC 12 participants inspired by the provided math content. The problem should encourage students to employ advanced logical reasoning and problem-solving strategies related to the given concepts. Craft it to be solvable in the AMC 12 format and difficulty level.

[Solution]: Present a comprehensive, step-by-step solution that both solves the problem and educates the student, around 250-350 words long. Clearly articulate the reasoning and methods used at each step, providing insight into the problem-solving process. Use language that challenges yet instructs high school students looking to improve their skills. Take care to format any equations properly using LaTeX or appropriate notation.

## Prompts for AIME-Level Problem

### Instruction

Inspired by the math content, create a math problem appropriate for advanced high school mathematics competitions like the AIME. The problem should challenge students in core areas tested on the AIME such as algebra, geometry, combinatorics, number theory and probability. Encourage creative problem-solving and deep mathematical thinking using pre-calculus level techniques. Avoid delving into college-level topics like abstract/linear algebra, topology, multivariable calculus, or advanced physics. Any physics concepts used should be kept at a basic mechanics level.

### Math Content

[Math Text Placeholder]

### Guidelines

[Problem]: Design an insightful and challenging problem suitable for AIME participants inspired by the math content. Focus on core AIME topics like algebra, geometry, combinatorics, number theory and probability. The problem should be solvable using clever applications of high school math and mathematical reasoning, without requiring knowledge of advanced college-level mathematics or physics. Aim for a difficulty level on par with actual AIME questions. [Solution]: Provide a detailed, step-by-step solution that would be enlightening to an advanced high school student, around 250-350 words long. Explain the reasoning process and techniques applied in each step. Use language that is mathematically sophisticated yet understandable to a well-prepared AIME participant. Justify any key insights or creative leaps in the solution process.

## I.2 Prompts for Tool Manipulation Data Synthesis

### Prompts for Tool Manipulation where Grade School or Secondary School Competition Level is Determined by Example

#### Instruction

Please gain inspiration from the following random math content to create a high-quality math problem and solve it with Python code. Present your output in two distinct sections: [Problem Description] and [Solution].

#### Math Content

[Math Text Placeholder]

#### Guidelines

1. [Problem Description]: This should be **completely self-contained**, providing all the contextual information one needs to understand and solve the problem.

2. [Solution]: Offer a comprehensive, **correct** solution that accurately addresses the [Problem Description] you provided using Python code. Summarize in natural language in the end and put the answer in /boxed.

#### Example

[Problem Description]

[Example Question Placeholder]

[/Problem Description]

[Solution]

[Example Python Solution Placeholder]

[/Solution]

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: According to the contributions proposed in Section 1, for the the first contribution, please refer to the Section 3 and Section 4.3, for the second contribution, please refer to the Section 3, and Appendix A, for the third contribution, please refer to Section 4.2.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Please refer to Appendix G.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We illustrate the necessary details of implementation, including details of prompting, training, and evaluation. Please refer to Appendix B, Appendix C, Appendix D, Appendix E, Appendix I

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We promise the code and the training data will be publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).



- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please refer to Appendix B, Appendix C, Appendix D, and Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our evaluation experiment is conducted under zero temperature for generation to avoid randomness. We report 18 different settings of evaluation datasets to demonstrate the performance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please refer to Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research is focused on LLMs for math reasoning, conforming with the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our research is focused on LLMs for math reasoning, which will not cause societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research is focused on designing LLMs for math reasoning which will not have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Please refer to Appendix D, Appendix C, Appendix E.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released code, data, and models are well documented including details about training, license, limitations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.