

# Exploring Mathematical Extrapolation of Large Language Models with Synthetic Data

**Haolong Li\***

Tongji University  
furlongli322@gmail.com

**Yu Ma**

Seed Foundation, ByteDance  
mayu.1231@bytedance.com

**Yinqi Zhang\***

East China Normal University  
zhang.inch@gmail.com

**Chen Ye<sup>†</sup>**

ESSC Lab, Tongji University  
yechen@tongji.edu.cn

**Jie Chen<sup>†‡</sup>**

Seed Foundation, ByteDance  
chenjiexjtu@gmail.com

## Abstract

Large Language Models (LLMs) have shown excellent performance in language understanding, text generation, code synthesis, and many other tasks, while they still struggle in complex multi-step reasoning problems, such as mathematical reasoning. In this paper, through a newly proposed arithmetical puzzle problem, we show that the model can perform well on multi-step reasoning tasks via fine-tuning on high-quality synthetic data. Experimental results with the open-llama-3B model on three different test datasets show that not only the model can reach a zero-shot pass@1 at 0.44 on the in-domain dataset, it also demonstrates certain generalization capabilities on the out-of-domain datasets. Specifically, this paper has designed two out-of-domain datasets in the form of extending the numerical range and the composing components of the arithmetical puzzle problem separately. The fine-tuned models have shown encouraging performance on these two far more difficult tasks with the zero-shot pass@1 at 0.33 and 0.35, respectively.

## 1 Introduction

Large Language Models (LLMs), as zero-shot and multi-task learners, have shown extraordinary capabilities across a variety of natural language tasks (Vaswani et al., 2017; Schulman et al., 2017; Radford et al., 2019; Ziegler et al., 2019; Brown et al., 2020; Kojima et al., 2022; Park et al., 2023; Chowdhery et al., 2023; Rafailov et al., 2024). However, even the most advanced LLMs face challenges when it comes to tackling complex multi-step reasoning problems, such as mathematical and scientific reasoning (Koncel-Kedziorski et al., 2016; Cobbe et al., 2021; Hendrycks et al., 2021; Wei et al., 2022; Chen et al., 2022; Gao et al., 2023;

Trinh et al., 2024). This comes from three main reasons: firstly, mathematical reasoning often requires quantitative multiple steps of deduction, since a single logical error is enough to derail a much larger solution (Lightman et al., 2023). Secondly, the lack of high-quality data limits LLMs’ ability to generalize and excel in mathematical reasoning tasks. Lastly, LLMs encounter difficulty in extrapolation, as they struggle to apply reasoning skills when solving unseen mathematical problems.

Many prior research has explored along these challenges. GPT-4 (Achiam et al., 2023), LLaMA (Touvron et al., 2023a,b), Gemini (Team et al., 2023), Minerva (Lewkowycz et al., 2022), Llemma (Azerbayev et al., 2023), Mistral (Jiang et al., 2023), WizardMath (Luo et al., 2023), MAMMOTH (Yue et al., 2023), ToRA (Gou et al., 2023) and Deepseek (Bi et al., 2024; Guo et al., 2024; Lu et al., 2024) have emerged as dominant models in popular mathematical reasoning benchmarks such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), CMATH (Wei et al., 2023) and AGIEval (Zhong et al., 2023). Moreover, process supervision and verifiers (Cobbe et al., 2021; Li et al., 2023; Uesato et al., 2022; Lightman et al., 2023; Yu et al., 2023) at the step level have also obtained widespread attention. However, mathematical extrapolation, particularly in terms of abstract forms, is often overlooked.

In this paper, we address the aforementioned challenges by introducing a novel and challenging arithmetical puzzle problem and making an initial attempt to solve them. Specifically, we propose a puzzle that needs multi-step calculations to generate a correct solution. Meanwhile, a data synthesis pipeline is developed to automatically generate a vast amount of high-quality data for supervised fine-tuning (SFT). And a series of LLMs based on open-llama-3B (Touvron et al., 2023a) are fine-tuned on this synthetic dataset. Furthermore, to demonstrate the reasoning abilities in extrapolation,

\*Work done during internship at ByteDance.

<sup>†</sup>Corresponding Author

<sup>‡</sup>Project Leader

Example of the Synthetic Data
—prompt— 34, 18, 31, 41, 19, 55: -110
—response— 31-34=-3, 19+41=60, 60/-3=-20, -20/18=-2, -2*55=-110

Table 1: Example of our synthetic data.

we have designed two out-of-domain benchmarks in the form of extending the numerical range and the composing components of the arithmetical puzzle problem. For the purpose of fair evaluation, we have restricted our models to greedy sampling in a zero-shot setting and provided a corresponding verifier. Our data scaling experiments demonstrate that as the amount of synthetic data grows, in-domain zero-shot pass@1 increases from 0.22 to 0.44, while the out-of-domain zero-shot pass@1 increases from 0.14/0.17 to 0.33/0.35.

Our major contributions can be concluded as: (1) We propose a novel arithmetical puzzle problem with corresponding data synthesis pipeline and out-of-domain benchmarks, to verify the multi-step reasoning and extrapolation capabilities of LLMs fine-tuned on synthetic data. (2) Experiments indicate that increasing the amount of high-quality synthetic data leads to performance enhancements across in-domain and out-of-domain datasets. (3) A comprehensive case study has been performed.

## 2 Problem Definition

### 2.1 Arithmetical Puzzle Problem

Arithmetical puzzle problem denotes a mathematical puzzle involving arithmetic operations and requires logical reasoning and numerical manipulation to derive a solution. The 24 Puzzle and Arithmetic Grid Puzzle are well-known examples of arithmetical puzzle problems.

In this paper, we propose a challenging arithmetical puzzle. Its objective is intricate yet precise: to deftly manipulate a set of given integers through a calculated sequence of arithmetic operations, to achieve a predetermined target integer. The problem strictly limits each integer to be used by one time exactly. For example, for the integers 3, 6, 7, 51, 58 and the target integer 4, one possible solution is:  $58-51=7$ ,  $6-7=-1$ ,  $3 \times -1=-3$ ,  $-3+7=4$ , as shown in Figure 5 in Appendix A.4.

### Algorithm 1 Data Synthesis Algorithm

---

```

1:  $S_{dataset}$  starts with an empty set
2: while  $size_{S_{dataset}} \leq size_{limit}$  do
3:   Sample  $\{X_i \mid 1 \leq i \leq N, X_i \sim U(1, V)\}$ 
4:    $L$  starts with an empty list
5:    $S \leftarrow \{X_i\}$ 
6:   for  $i = 1$  to  $N - 1$  do
7:     Randomly select  $a_i, b_i \in S$ 
8:     Randomly select  $ops_i \in \{+, -, \times, \div\}$ 
9:      $c_i \leftarrow a_i \ ops_i \ b_i$ 
10:     $S \leftarrow S - \{a_i\} - \{b_i\}$ 
11:     $S \leftarrow S \cup \{c_i\}$ 
12:     $L \leftarrow L + \{a_i.ops_i.b_i, c_i\}$ 
13:  end for
14:   $T \leftarrow c_{N-1}$ 
15:  if  $\{L, T\} \notin S_{dataset}$  then
16:     $S_{dataset} \leftarrow S_{dataset} \cup \{L, T\}$ 
17:  end if
18: end while

```

---

### 2.2 Data Synthesizing

Given the arithmetical puzzle described above in Section 2.1, we create a data synthesizing pipeline to efficiently generate the proposed dataset.

Denote the set of candidate integers as  $X = \{X_1, X_2, \dots, X_N\}$  and the target number as  $T$ , where  $N$  is the total number of candidate integers in a puzzle sample. Each candidate integer  $X_i$  is independently sampled from a uniform distribution  $X_i \sim U(1, V)$ , where  $V$  is the upper bound of the sampled integers. To avoid data overlapping, we have strictly ensured that for each puzzle, the candidate integers are a set of distinct numbers. The arithmetic operators involved in this problem are  $ops = \{+, -, \times, \div\}$  and all operations are limited to integer operations. For example, when solving the puzzle with a division operator, the operation should be considered in integer division like  $14/3 = 4$ . The detailed steps of synthesizing data for this puzzle is described in Algorithm 1.

Besides, to construct the SFT dataset, the prompt is deliberately designed to exclude any natural language cues and instead focuses on purely symbolic language. See Table 1 for an example of the constructed prompt and response.

### 2.3 Dataset

We split the dataset into training and in-distribution and out-of-distribution test dataset by controlling the total number of candidate integers  $N$  and the upper bound of the sampled integers  $V$ . We set

---

**Algorithm 2** Verifier Algorithm

---

```
1:  $\{X_i \mid 1 \leq i \leq N\} \leftarrow X_{prompt}$ 
2:  $T \leftarrow T_{prompt}$ 
3:  $Eqs \leftarrow Solution_{generated}$ 
4:  $S \leftarrow \{X_i\}$ 
5:  $Flag_{verifier} \leftarrow False$ 
6: for  $eq_i \in Eqs$  do
7:   if  $eq_i$  is a legal equation then
8:      $a_i, ops_i, b_i, c_i \leftarrow ParseEq(eq_i)$ 
9:     if  $a_i, b_i \in S$  then
10:       $S \leftarrow S - \{a_i\} - \{b_i\}$ 
11:       $S \leftarrow S \cup \{c_i\}$ 
12:     else
13:       break
14:     end if
15:   else
16:     break
17:   end if
18: end for
19: if  $c_{N-1} = T$  then
20:    $Flag_{verifier} \leftarrow True$ 
21: end if
```

---

$V = 60$  for the training dataset, and sampled the candidate integers with  $N = 5, 6, 7$ . Three training datasets with different sizes scaling from 1 million to 10 millions and 100 millions are generated. And another 7500 samples (2500 samples for each  $N$ ) under the same setting are generated as the in-distribution test dataset. Figure. 1 shows the distribution of  $N$  and  $X$  in these three training datasets. And the corresponding distribution of the tokenized prompt and response length is shown in Figure. 2.

To further evaluate the model’s performance on extrapolation, we have also designed two benchmarks of out-of-distribution dataset:

**Numerical OOD test datasets.** The upper bound of the sampled integers  $V$  is raised to 100 and 1000 separately to test the model’s generalization ability with unseen larger numbers. Specifically, 6000 samples are generated for each value of  $V$  with 2000 samples for each  $N$ . An additional filtering pipeline is applied to ensure that for each sample, there exists at least one integer  $X_i$  that satisfies  $60 < X_i < 100$  for the dataset with  $V = 100$  and  $100 < X_i < 1000$  for that with  $V = 1000$ .

**Form OOD test dataset.** In mathematics, abstract forms often extend, such as expanding from a two-variable linear equation to one with three variables. For the proposed arithmetic puzzle, the

extrapolation of abstract forms can be achieved by changing the number of candidate integers  $N$ . Clearly, when  $N$  increases, the exploration space leading to a feasible solution would expand exponentially, which results in an increased demand for precise reasoning steps. From another perspective, when the total number of the candidate integers changes, it actually requires the model’s ability to absorb and adapt to the puzzle’s abstract forms. Therefore, to test the model’s generalization capability from this point of view, we create another benchmark for OOD test dataset with 5000 samples generated with setting  $N$  to 8. To control variables, all the candidate integers in this dataset are sampled with the same upper bound  $V = 60$  as the training dataset.

### 3 Model

#### 3.1 Framework

We adopt the llama architecture (Touvron et al., 2023a) and employ low-rank adaptation (LoRA) tuning (Hu et al., 2021) based on the implementation of TRL full stack library (von Werra et al., 2020). LoRA achieves a remarkable reduction of 89% in our trainable parameters, from 3B to 0.3B.

#### 3.2 Implementation Details

We train our model by fine-tuning open-llama-3B. We systematically apply left-padding to the query text and right-padding to the answer text to control the overall context length. All experiments are conducted with  $8 \times$  NVIDIA A100-SXM4-80GB GPUs. The specific hyperparameter settings are listed in Table 3 in Appendix A.1.

## 4 Experiments

### 4.1 Evaluation

For the fine-tuned model, we use the greedy decoding strategy in a zero-shot setting to generate responses. To measure the model’s performance on the proposed puzzle, a corresponding verifier is designed to automatically evaluate the correctness of the responses. Specifically, a solution is deemed correct if it satisfies the following rules:

- No extra or illegal characters.
- There are only  $N - 1$  equations and all the corresponding calculations are correct.
- $F(X_1, \dots, X_N \mid ops) = T$ .
- All  $\{X_i \mid i \in \{1, 2, \dots, N\}\}$  and the intermediate calculation results are only used once.

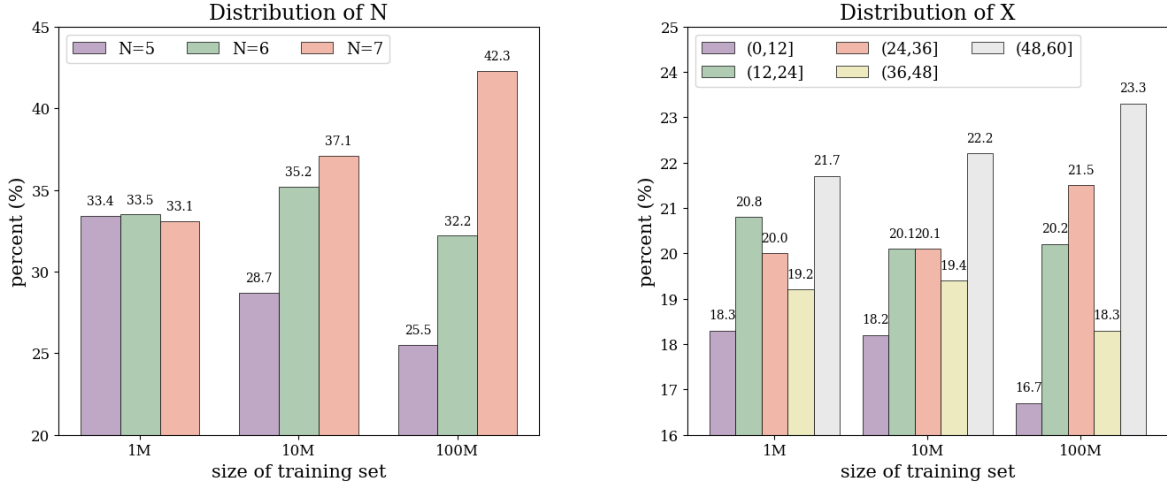


Figure 1: Distributions of  $N$  and  $X$  for different training set sizes (1M / 10M / 100M samples).  $N$  denotes the total number of candidate integers of our puzzle,  $X = (X_1, X_2, \dots, X_N)$  denotes the candidate integers.

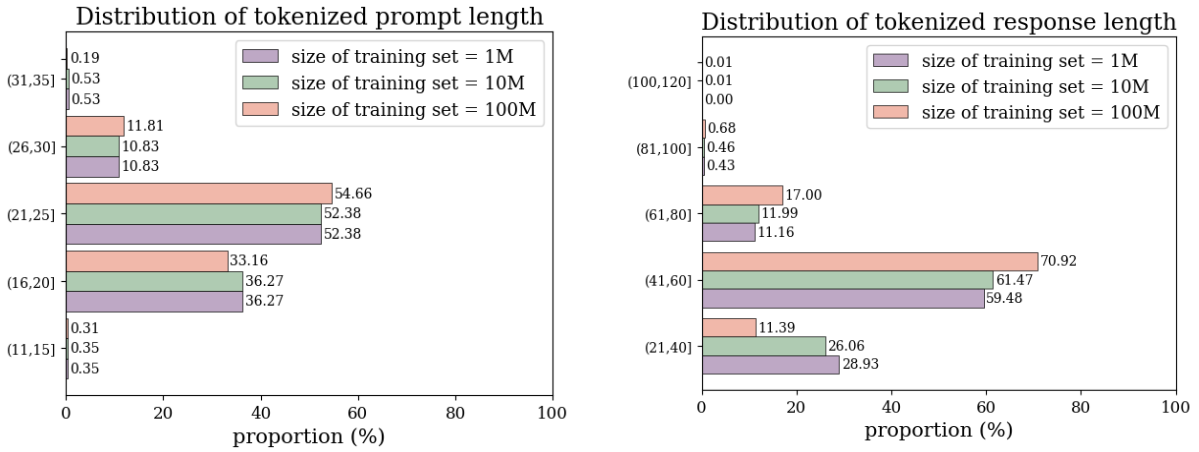


Figure 2: Distributions of the tokenized prompt and response lengths for different training set sizes (1M / 10M / 100M samples).

The detailed steps of evaluating the solution for this puzzle is described in Algorithm 2.

## 4.2 Results

As mentioned in Section 2.3, we have generated three training datasets with different sizes to explore the data scaling effects on the fine-tuned model. The pass@1 rate on different in-distribution and out-of-distribution test datasets are shown in Table 2. When the model is fine-tuned with 100M samples, it achieves the highest score with a zero-shot pass@1 of 0.44 in the in-distribution test dataset, and 0.33 and 0.35 in the two OOD datasets, respectively.

Furthermore, we have shown the training curves of the model fine-tuned on these three datasets in Figure 3. From Figure 3, a faster decaying rate is clearly observed in the training loss when increasing the training data size, which is consistent with

the rapid increase of the pass@1 rate evaluated on the in-distribution dataset. The same enhancement of the performance also occurs in the two OOD test datasets as shown in Table 2.

Additionally, we have also conducted tests of this puzzle on the base model (open-llama-3B) and several other open-source and closed-source models with both few-shot and CoT prompting. The results and some of the generated cases are shown in Appendix A.2, demonstrating the necessity of fine-tuning with regard to solving such puzzle problems.

## 4.3 Case Studies

We further demonstrate the different solutions provided by models trained with 1M / 10M / 100M training data on the form OOD test dataset for several challenging queries. As shown in Figure 4 in Appendix A.3, the model trained on 1M sam-

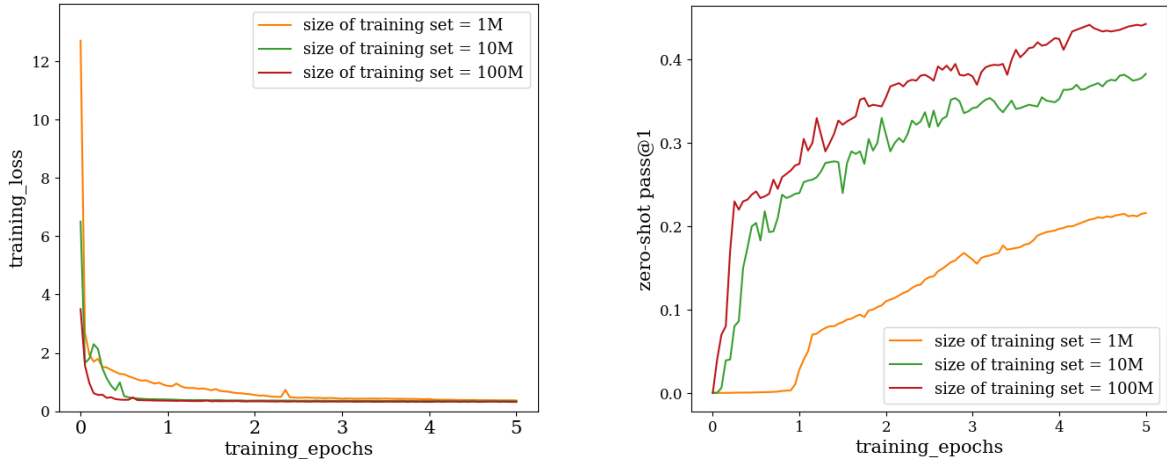


Figure 3: The training loss and zero-shot pass@1 on ID dataset for different training set sizes (1M / 10M / 100M samples).

Dataset	Range	Number of Integers	Fine-tuned on 1M	Fine-tuned on 10M	Fine-tuned on 100M
ID	[1,60]	5	0.224	0.428	<b>0.471</b>
		6	0.208	0.363	<b>0.432</b>
		7	0.205	0.360	<b>0.425</b>
Total ID	[1,60]	5,6,7	0.216	0.383	<b>0.443</b>
Numerical OOD	[1,100]	5	0.163	0.239	<b>0.364</b>
		6	0.137	0.199	<b>0.331</b>
		7	0.126	0.186	<b>0.315</b>
Total Numerical OOD	[1,100]	5,6,7	0.141	0.205	<b>0.326</b>
Numerical OOD	[1,1000]	5	0.131	0.181	<b>0.229</b>
		6	0.030	0.051	<b>0.063</b>
		7	0.111	0.163	<b>0.220</b>
Total Numerical OOD	[1,1000]	5,6,7	0.091	0.132	<b>0.170</b>
Form OOD	[1,60]	8	0.169	0.231	<b>0.352</b>

Table 2: Zero-shot pass@1 of the model fine-tuned with different training set sizes (1M / 10M / 100M samples) on ID, numerical OOD, and form OOD test datasets. The best results are **highlighted**.

ples is still limited to a fixed number of reasoning steps, whereas the models trained on 10M / 100M samples exhibit a higher-level understanding of the problem and perform an adequate number of reasoning steps. However, compared to the model trained on 100M samples, the model trained on 10M samples may still encounter computational or logical errors in the final step of reasoning.

## 5 Conclusion

Large language models (LLMs) are intrinsically zero-shot and multi-task learners. However, mathematical reasoning still poses challenges for LLMs, we propose that the reasons can be mainly categorized into three folds: (1) Requirement of multi-step derivation; (2) Lack of high quality data for fine-tuning; (3) Difficulty in extrapolation. In this paper, we design an arithmetical puzzle and make an early attempt to solve these challenges. We develop a 24-point puzzle-like problem which asks

for multi-step calculations to arrive at the correct answer. A corresponding data synthesis pipeline is proposed to generate an arbitrary amount of high-quality data, on which a series of LLMs are fine-tuned. In order to verify the extrapolation capability of our models, we have designed two out-of-domain benchmarks and show that our model achieves competitive performance. Furthermore, a data scaling experiment is conducted and it is concluded that by increasing the amount of training data, both the training loss and in/out-of-domain performance of the fine-tuned model improve accordingly.

## Acknowledgements

We appreciate Peng Sun for providing the initial SFT dataset, and Xintian Han for suggestions about the reward calculation and ablation study. We would also like to thank Liang Xiang and Xun Zhou for the helpful discussions across the project.

## 6 Limitations

In this study, we have explored the mathematical extrapolation of Large Language Models (LLMs) and discovered that, with high-quality synthetic data, LLMs demonstrates certain generalization capabilities in mathematical extrapolation. However, LLMs have not yet fully mastered this capability, and it remains uncertain if this ability can be extended to other complex mathematical tasks. In the future, our research will focus on investigating and enhancing this capability, aiming to empower LLMs to explore unsolved mathematical problems through leveraging our existing knowledge.

## 7 Ethics Statement

In this research, we adhere to strict ethical guidelines and principles. The study has been designed and implemented with respect for the rights, privacy, and well-being of all individuals involved. All of our data is synthesized using our proposed data synthesis algorithm, ensuring compliance with relevant regulations and standards. Our findings and conclusions are reported accurately and objectively, avoiding any misrepresentation or manipulation of data. The entire process and outcomes are free from intellectual property and ethical legal disputes.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.
- Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujia Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. Mawps: A math word problem repository. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157.

- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.
- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making large language models better reasoners with step-aware verifier. *arXiv preprint arXiv:2206.02336*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. Cmath: can your language model pass chinese elementary school math test? *arXiv preprint arXiv:2306.16636*.
- Fei Yu, Anningzhe Gao, and Benyou Wang. 2023. Outcome-supervised verifiers for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

## A Appendix

### A.1 Hyperparameter Settings

In the SFT stage, we follow common fine-tuning hyperparameter settings for our model. We set learning rate to  $1e-4$  and adopt the cosine learning rate scheduler. We use low-rank adaptation (LoRA) tuning with a rank of 5,  $\alpha$  of 32, and dropout of 0.05. And we employ Adamw optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$  and  $\epsilon = 1e-9$ . Eight NVIDIA A100-SXM4-80GB GPUs are used to train the model with a batch size of 50 and the maximum epoch set to 5. Detailed settings are listed in Table 3.

Hyperparameter	Value	Hyperparameter	Value
Learning Rate	$1e-4$	Epochs	5
Learning Rate Scheduler	Cosine	Optimizer	Adamw
Warmup Step	0	Optimizer $\beta_1$	0.9
GPU Num	8	Optimizer $\beta_2$	0.95
Batch Size Per GPU	50	Optimizer $\epsilon$	$1e-9$
Max Query Length	36	Precision	AMP
Max Answer Length	130	LoRA Rank	8
Max Generated Length	167	LoRA $\alpha$	32
Precision	bfloat16	LoRA Dropout	0.05
Vocabulary Size	32002	Seed	1234

Table 3: Hyperparameter Settings.

### A.2 Evaluation of the Base Model

We evaluate the base model (open-llama-3B) on the proposed arithmetical puzzle problem. As shown in Table 4 and Table 5, with either the few-shot prompting (2-Shot, 8-Shot) or Chain-of-Thought (CoT), the base model performs poorly on the puzzle. We propose this is due to the symbolic form of our prompt, the model needs to understand the underlying pattern in order to solve the arithmetical puzzle. Without fine-tuning on the synthetic data, the model may struggle to comprehend such type of prompt.

Approach	ID	OOD-N	OOD-F	OOD-1000
2-Shot Prompting	0	0	0	0
8-Shot Prompting	0	0	0	0
Chain-of-Thought	0	0	0	0

Table 4: Evaluation of the base model with few-shot and Chain-of-Thought prompting. As expected, the base model performs poorly across all the prompting techniques.

Chain-of-Thought Prompting	Generated Response of the Base Model
Let's play an arithmetic puzzle game. First, we are given a set of numbers: 8, 37, 11, 26, 15. The objective is to use operations like addition, subtraction, multiplication, and division to obtain the number 8, using each number exactly once. Please provide the reasoning process step by step.	The first thing to do is to find the number of ways to get 8. There are 4 ways to get 8 by adding 8 to 37, 11, 26, 15. There are 4 ways to get 8 by subtracting 8 from 37, 11, 26, 15. There are 4 ways to get 8 by multiplying 8 from 37, 11, 26, 15. There are 4 ways to get 8 by dividing 8 from 37, 11, 26, 15. There are 4 ways to get 8 by adding 8 from 37, 11, 26, 15. There are 4 ways to get 8 by subtracting 8 from 37, 11, 26, 15. There are 4 ways to get 8 by multiplying 8 from 37, 11, 26, 15. There are 4 ways to get 8 by dividing 8 from 37, 11, 26, 15. ...

Table 5: An example of Chain-of-Thought prompting and the generated response of the base model.



We further test several open-source (Llama-2-7B (Touvron et al., 2023a), Deepseek-Coder-33B (Guo et al., 2024)) and closed-source models (GPT4 (Achiam et al., 2023)) with few-shot prompting. As shown in Table 6, these models also perform poorly on our benchmarks. In Table 7, we provide an example of the CoT prompting and the generated responses from these models.

Model	ID	OOD-N	OOD-F	OOD-1000
GPT4	0.002	0.001	0	0.001
Llama-2-7B	0	0	0	0
Deepseek-Coder-33B	0.001	0	0.001	0

Table 6: Evaluation results of Llama-2-7B, Deepseek-Coder-33B, and GPT4 on our proposed benchmarks.

Few-Shot Prompting	Model	Response
<p>A puzzle problem. The first two lines are examples. Please solve the third line:            36, 32, 57, 55, 11:30S11/36=0,            0+32=32, 55+32=87, 87-57=30            17, 6, 20, 48, 30:-40S17+48=65,            30/6=5, 20+5=25, 25-65=-40            8, 37, 11, 26, 15:8S</p>	GPT4	<p>Sure, I'll solve the third line of the puzzle using the pattern established in the first two lines:            Firstly, divide the fifth number by the first number, which will yield an integer part:  <math>15 / 8 = 1</math> (ignore the decimal part)            Next, add the result to the second number:  <math>1 + 37 = 38</math>            After that, add this result to the fourth number:  <math>38 + 26 = 64</math>            Final step, subtract the third number from this result:  <math>64 - 11 = 53</math>            However, the expected output provided is 8, but following the pattern above, the answer seems to be 53. It's possible that there are some other rules or factors at play, or there could be an error in the input or expected output.</p>
	Llama-2-7B	<p>Sure, I'd be happy to help you solve the third line of the puzzle!            The equation is: 8,37,11,26,15:8S            To solve this equation, we can start by simplifying the fraction:  <math>8/8 = 1</math>            So, the solution to the equation is:  <math>11,26,15 = 52</math>            I hope this helps! Let me know if you have any other questions.</p>
	DeepSeek-Coder-33B	<p>The third line of the puzzle problem is:            17, 6, 20, 48, 30:-40S17+48=65, 30/6=5, 20+5=25, 25-65=-40            The operations and numbers are:            1. <math>17 + 48 = 65</math>            2. <math>30 / 6 = 5</math>            3. <math>20 + 5 = 25</math>            4. <math>25 - 65 = -40</math>            Let's solve these operations:            1. <math>17 + 48 = 65</math>            2. <math>30 / 6 = 5</math>            3. <math>20 + 5 = 25</math>            4. <math>25 - 65 = -40</math>            So, the solution to the third line is:  <math>17 + 48 = 65</math>  <math>30 / 6 = 5</math>  <math>20 + 5 = 25</math>  <math>25 - 65 = -40</math>            This line of the puzzle problem is solved.</p>

Table 7: An example of few-shot prompting and the generated responses of GPT4, Llama-2-7B, and Deepseek-Coder-33B. We provide the models with two examples before the puzzle. As shown, all of the models fail to solve the given problem. GPT4 seems to understand the requirement of the puzzle, while the other two fail.

As shown in Table 7, Llama-2-7B fails to understand the requirement of the puzzle and just outputs two meaningless equations. Deepseek-Coder-33B treats the second example in few-shot prompting as the puzzle, and repeats the same calculations three times. It seems that GPT4 has well understood the prompt and used all the candidate integers only once, the calculations within the generated response are all right, while the solution is wrong. Actually, such kind of problem is very challenging, as the model needs to infer the requirement of the puzzle from the provided examples and then figure out the correct solution.

### A.3 Case Study

Prompt: 21, 55, 40, 54, 17, 15, 20, 60:20

Response 1M(✗):  $55/15=3$ ,  $21/54=0$ ,  $40-17=23$ ,  $23-3=20$ ,  $0*60=0$ ,  $20-0=20$  ?

Response 10M(✓):  $55/17=3$ ,  $21-3=18$ ,  $15/54=0$ ,  $18-60=-42$ ,  $0*-42=0$ ,  $0/40=0$ ,  $0+20=20$

Response 100M(✓):  $15/55=0$ ,  $21/54=0$ ,  $0*40=0$ ,  $0*60=0$ ,  $0*17=0$ ,  $0+20=20$ ,  $20-0=20$

Prompt: 28, 48, 1, 15, 45, 20, 57, 47:0

Response 1M(✗):  $45/48=0$ ,  $15*20=300$ ,  $47*0=0$ ,  $28/57=0$ ,  $0*300=0$ ,  $0-0=0$  ?

Response 10M(✓):  $45/1=45$ ,  $48/20=2$ ,  $28/47=0$ ,  $15/2=7$ ,  $57-45=12$ ,  $12-7=5$ ,  $0*5=0$

Response 100M(✓):  $47/15=3$ ,  $20/48=0$ ,  $3/57=0$ ,  $0*28=0$ ,  $0*1=0$ ,  $0*45=0$ ,  $0+0=0$

Prompt: 46, 3, 49, 27, 39, 20, 53, 58:0

Response 1M(✗):  $49/58=0$ ,  $20*3=60$ ,  $27/46=0$ ,  $0*53=0$ ,  $0*39=0$ ,  $0-0=0$  ?

Response 10M(✓):  $39/49=0$ ,  $20/53=0$ ,  $0*58=0$ ,  $0*3=0$ ,  $0*46=0$ ,  $27*0=0$ ,  $0-0=0$

Response 100M(✓):  $20/53=0$ ,  $39/49=0$ ,  $0*27=0$ ,  $0*3=0$ ,  $0*46=0$ ,  $0*58=0$ ,  $0+0=0$

Prompt: 53, 55, 47, 5, 6, 7, 22, 22:55

Response 1M(✗):  $53-22=31$ ,  $5/47=0$ ,  $22/6=3$ ,  $31*0=0$ ,  $0/7=0$ ,  $55-0=55$  ?

Response 10M(✗):  $53/22=2$ ,  $22/55=0$ ,  $7-47=-40$ ,  $0+-40=-40$ ,  $6-2=4$ ,  $4-5=1$ ,  $1--40=51$

Response 100M(✓):  $22/5=4$ ,  $47/53=0$ ,  $4/7=0$ ,  $0*22=0$ ,  $0*6=0$ ,  $0+0=0$ ,  $55-0=55$

Prompt: 2, 37, 22, 41, 56, 19, 60, 40:58

Response 1M(✗):  $41-22=19$ ,  $19+40=59$ ,  $59-37=22$ ,  $22-19=3$ ,  $56/3=18$ ,  $18*2=58$  ?

Response 10M(✗):  $22/41=0$ ,  $0/2=0$ ,  $19/60=0$ ,  $0*0=0$ ,  $0*40=0$ ,  $0+37=37$ ,  $37+56=85$

Response 100M(✓):  $41/22=1$ ,  $19/60=0$ ,  $0+56=56$ ,  $37/2=18$ ,  $18/1=18$ ,  $40/18=2$ ,  $2+56=58$

Prompt: 18, 16, 48, 12, 57, 7, 56, 32:-87

Response 1M(✗):  $12-56=-44$ ,  $18+-44=-26$ ,  $57+32=89$ ,  $16+-26=-10$ ,  $48/-10=-5$ ,  $-5-89=-87$  ?

Response 10M(✗):  $18-56=-38$ ,  $12+32=44$ ,  $48+-38=10$ ,  $10-44=-34$ ,  $16+-34=-18$ ,  $-18-57=-75$ ,  $-75-7=-87$

Response 100M(✓):  $16-56=-40$ ,  $12+57=69$ ,  $18+69=87$ ,  $32+7=39$ ,  $39/48=0$ ,  $0*-40=0$ ,  $0-87=-87$

Prompt: 18, 3, 33, 51, 28, 18, 59, 4:-79

Response 1M(✗):  $18-59=-41$ ,  $33+4=37$ ,  $37+51=88$ ,  $18+-41=-23$ ,  $28-88=-60$ ,  $-23+-60=-79$  ?

Response 10M(✗):  $18-3=15$ ,  $15-51=-36$ ,  $33+18=51$ ,  $59+51=110$ ,  $110-28=82$ ,  $4+82=86$ ,  $-36-86=-170$

Response 100M(✓):  $18-59=-41$ ,  $33+18=51$ ,  $51+51=102$ ,  $28-102=-74$ ,  $3/-41=-1$ ,  $-74+-1=-75$ ,  $-75-4=-79$

Figure 4: Cases from the form OOD test dataset. The correct steps are highlighted in green, while the incorrect steps in red. Generally speaking, performance of model fine-tuned with 1M training data is the worst.

A.4 Visualization of the Proposed Puzzle

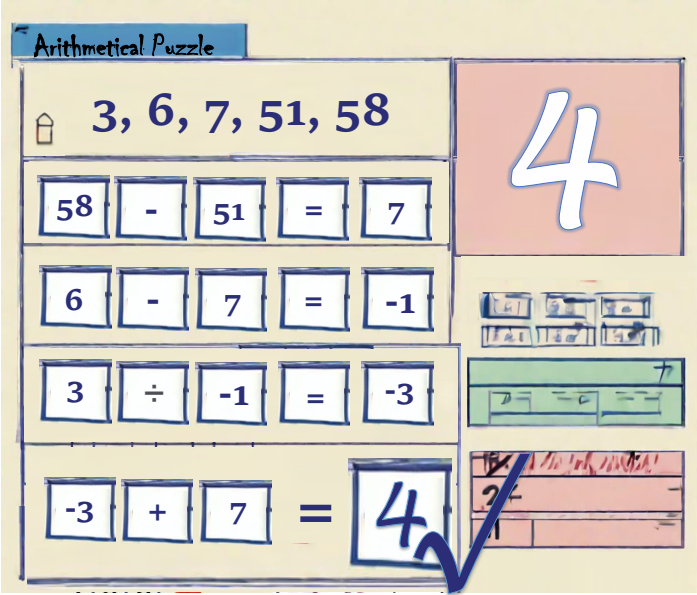


Figure 5: Visualization of the proposed arithmetical puzzle. Given the candidate integers 3, 6, 7, 51, 58 and the target integer 4, the answer is  $58 - 51 = 7, 6 - 7 = -1, 3 \times (-1) = -3, -3 + 7 = 4$ .