Rationales for Answers to Simple Math Word Problems Confuse Large Language Models

Yidan Zhang[♠] **Mingfeng Xue**[♠] **Dayiheng Liu**[♠] **Zhenan He**[♠*]
[♠]College of Computer Science, Sichuan University
zhenan@scu.edu.cn

Abstract

Recently, large language models (LLMs) have demonstrated breakthrough mathematical problem-solving capabilities in grade school math word problems (MWP). For example, on the MWP benchmark GSM8K, the accuracy of GPT-3.5-Turbo and MetaMath-70B reaches 80.80% and 82.30%, respectively. One question arises, does it mean that LLMs have truly mastered related mathematical problem-solving abilities? In this paper, by presenting two types of benchmarks, where MCGSM8K aims at selecting one correct solution from four solutions, while GSM8K-Judgement judges whether a solution to a given question is true or false, we demonstrate that the ability of most LLMs to evaluate the mathematical reasoning process of MWP is far from sufficient. To compensate for this issue, we propose hybrid supervised fine-tuning data from the training data of GSM8K, MCGSM8K, and GSM8K-Judgement, which significantly improves performance on the proposed reasoning process evaluation benchmarks. For example, fine-tuning improves the performance of LLaMA-2-13B from 33.51% to 70.89% on MCGSM8K. In conclusion, we experimentally demonstrate that most LLMs have limited ability to evaluate the mathematical reasoning process of MWP, which can be enhanced through fine-tuning.

1 Introduction

It is reported that general close-source large language models (LLMs) have demonstrated promising performance on several mathematical word problems (MWP) benchmarks, e.g., GPT-4 (OpenAI, 2023) and GPT-3.5-Turbo (OpenAI, 2022) achieving the accuracy of 92.00% and 80.80% on grade school MWP benchmark GSM8K (Cobbe et al., 2021a), respectively. With the development of prompt-based methods (Fu et al., 2023; Wang et al., 2023a) and finetuning-based methods (Yu et al., 2023; Yue et al., 2023; Yuan et al., 2023), mathematical specialized LLMs tuned on specific tasks also exhibit competitive performance. For example, MetaMath-70B (Yu et al., 2023), WizardMath-70B (Luo et al., 2023), and MAmmoTH-70B (Yue et al., 2023) achieves 82.30%, 81.60%, and 76.90% on GSM8K, respectively. Now, one question arises, does the excellent performance demonstrate that these LLMs truly master related mathematical problem-solving abilities, such as the ability to evaluate the mathematical reasoning process of MWP?

Intuitively, picking one correct solution from possible solutions is easy for humans, as it just requires evaluating the correctness of the reasoning process. In comparison, reasoning the answer based on the open-formed question is difficult, which requires analyzing the problem, step-by-step reasoning, and deriving the final result (Cobbe et al., 2021b). Building upon this premise, we design a simple mathematical reasoning processes evaluation benchmark, MCGSM8K [1] aiming at choosing one correct solution from four options, as shown in Figure 2. Then, we utilize a few-shot (Chen et al., 2022b; Min et al., 2022) setup to test the performance of typical general open-source models LLaMA-2, general closed-source models GPT-3.5-Turbo and GPT-4, as well as mathematical specialized models MetaMath and MAmmoTH on it. However, our experimental results in Figure 1 and Table 1 reveal that most LLMs lag far behind on MCGSM8K. For example, the accuracy of LLaMA-2-70B, MetaMath-70B, and GPT-3.5-Turbo drop from 56.80% to 38.29%, 82.30% to 34.87%, and 80.80% to 40.56%, respectively. Specifically, each solution of MCGSM8K contains both the final answer and a step-by-step reasoning process (rationale). We collect incorrect solutions by regenerating solutions for each question in the test set of GSM8K. To keep the quality and di-

* Corresponding author.

[1]MCGSM8K is publicly available at https://github.com/SCUZPP/MCGSM8K.git

Findings of the Association for Computational Linguistics ACL 2024, pages 8853–8869
August 11-16, 2024

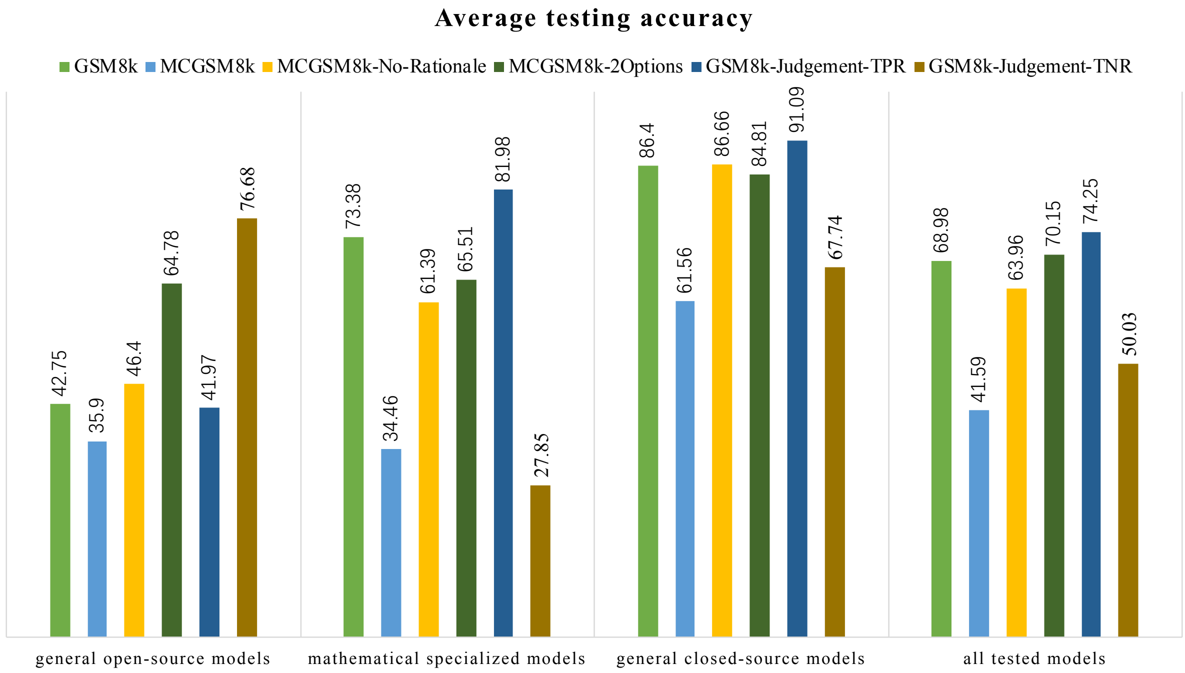

Figure 1: Average few-shot testing accuracy by general open-source models (LLaMA-2 with 13B and 70B), mathematical specialized models (MetaMath with 13B and 70B, and MAmmoTH with 13B and 70B), general closed-source models (GPT-3.5-Turbo and GPT-4), and all tested models.

versity of MCGSM8K, we use multiple advanced LLMs, e.g., Qwen (Bai et al., 2023), LLaMA-2 (Touvron et al., 2023b), MetaMath (Yu et al., 2023), and WizardLM (Xu et al., 2023), through few-shot Chain-of-thought (CoT) (Wei et al., 2022) prompting to generate incorrect solutions.

To comprehensively investigate the performance of LLMs on MCGSM8K, we progressively conduct the following experiments with a few-shot setting on three well-designed benchmarks as shown in Figure 2. First of all, the main reason for the poor performance of LLMs on MCGSM8K might include a) the model's inability to solve multiple-choice format questions, and b) the model's inability to evaluate the reasoning process. To verify, we propose a conventional multiple-choice-question benchmark MCGSM8K-No-Rationale by removing the rationale and leaving only the final answer for each option from MCGSM8K. The average accuracy of all tested models on MCGSM8K-No-Rationale (63.96%) is 25.50% higher than that on MCGSM8K (41.59%) and close to that on GSM8K (68.98%). The result reveals that the poor performance might be due to the model has difficulty in evaluating the reasoning process instead of the answer directly. Furthermore, we analyze the performance of LLMs by reducing the difficulty of solving the problem in MCGSM8K to half. Specifically, we remove any two incorrect options for each question in MCGSM8K, resulting in MCGSM8K-2Options aiming at selecting one correct solution from two. However, the average accuracy of all tested models on the two-choice-question benchmark MCGSM8K-2Options is merely 1.17% higher than that on the open-formed-question benchmark MCGSM8K (70.15% vs. 68.98%). MetaMath-70B and LLaMA-2-70B achieve an accuracy of 66.94% and 64.22%, which is merely 16.94% and 14.22% higher than the random-chance accuracy of 50%, respectively. This verifies the ability of most LLMs to evaluate the mathematical reasoning process is insufficient. Are LLMs insufficient to evaluate the correct solution or the incorrect solution? To figure this out, we propose a true-or-false-question benchmark GSM8K-Judgement to directly judge the correctness of the solution to a given question. In general, it is easier for humans to identify incorrect solutions than correct solutions, as the former only requires identifying a certain incorrect step, while the latter must ensure the correctness of all steps. Experimental results on general open-source models show that the average True Negative rate (TNR) is significantly better than the average True Positive rate (TPR), which is consistent with human behavior. However, when it comes to mathematical specialized models and GPT-3.5-Turbo, the situation is completely reversed.

Through the above experimental analysis, it can be concluded that most LLMs have a poor ability to evaluate the mathematical reasoning process of MWP. Mathematical specialized models are mainly fine-tuned on abundant correct solutions, which greatly improves the ability to identify correct solutions while causing catastrophic forgetting in identifying incorrect solutions. In addition, we hypothesize that most LLMs only catch spurious signals in specific datasets resulting in "solving" the datasets while not mastering abilities related to mathematical problem-solving.

In this paper, we try to compensate for these shortcomings by finetuning models on hybrid training samples from GSM8K, MCGSM8K, and GSM8K-Judgement. This contributes to enhancing the generalization ability to solve mathematical problems, mastering new data distribution on the MCGSM8K and GSM8K-Judgement, identifying correct solutions, as well as learning to analyze and evaluate the reasoning process. Specifically, we use LLaMA-2-13B as the base model. After fine-tuning, we observe a substantial improvement in accuracy with an increase of +37.38% on MCGSM8K, +41.24% on TPR, +8.87% on TNR, +7.43% on MCGSM8K-No-Rationale, and +16.62% on GSM8K. The result demonstrates that fine-tuning can greatly improve the mathematical reasoning process evaluation ability of LLMs.

Our main contributions can be summarised as follows:

- To explore whether LLMs have mastered the ability to evaluate the mathematical reasoning process of MWP, we carefully create two kinds of benchmarks. The first category is multiple-choice questions, including MCGSM8K aiming at choosing one correct solution from four options, MCGSM8K-2Options containing only two options, and MCGSM8K-No-Rationale with only the final answer in each option. The second type is true or false questions, including GSM8K-Judgement to judge the correctness of the solution to a given question.

- We conduct experiments with typical general open-source models, general closed-source models, and mathematical specialized models on the four benchmarks. The experimental results reveal that existing LLMs except GPT-4 have a poor ability to evaluate the mathematical reasoning process of MWP. Meanwhile, fine-tuning with only correct solutions improves the performance in evaluating correct solutions, but leads to a huge performance drop in evaluating incorrect solutions. Furthermore, we experimentally demonstrate that these drawbacks of LLMs can be alleviated to a certain extent through fine-tuning.

2 Related Work

Large Language Models. LLMs with billions of parameters trained on extensive large-scale corpora of textual data have led to massive changes in the field of AI. OpenAI's GPT series (Brown et al., 2020; OpenAI, 2023), which is the most representative general closed-source LLM, opens the era of pre-trained LLMs, where a large number of prominent instances are launched one after another involving Anthropic's Claude 2 (Bai et al., 2022), Google's PaLM (Chowdhery et al., 2023; Anil et al., 2023), DeepMind's Chinchilla (Hoffmann et al., 2022), and Gopher (Rae et al., 2021). Subsequently, numerous general open-source LLMs have been released, whose code and weight parameters are open to the public. Typical examples include LLaMA (Touvron et al., 2023a,b), GLM-130B (Zeng et al., 2023), OPT (Zhang et al., 2022), Falcon (Penedo et al., 2023), and so on. Although general closed-source LLMs, e.g., GPT-3.5, GPT-4, and PaLM-2, have achieved considerable advancements in several MVP tasks such as GSM8K and NumGLUE (Mishra et al., 2022), the performance of general open-source LLMs is far from satisfactory.

Large Language Models for Mathematical Reasoning. Chain-of-thought (CoT) (Wei et al., 2022) reasoning by designing better prompts has been proposed to generate step-by-step solutions leading to improved performance in complex reasoning. To improve the mathematical reasoning capabilities of general open-source LLMs, existing methods focus on CoT prompting for augmenting fine-tuning data. WizardMath (Luo et al., 2023) utilizes few-shot CoT prompting to re-generate solutions for GSM8K and MATH (Hendrycks et al., 2021), then uses this data to construct SFT data. MetaMath (Yu et al., 2023) aims to improve finetuning-based methods by answer augmentation and available mathematical questions bootstrap to

GSM8K
Question: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Rationale: Weng earns 12/60 = $<<12/60=0.2>>0.2 per minute.
Working 50 minutes, she earned 0.2*50 = $<<0.2*50=10>>10.
Final Answer: 10

MCGSM8K
Question: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Options:
(A) 600. Rationale: Weng earns $12*50 = $600 for 50 minutes of babysitting.
(B) 2430. Rationale: Weng earns $12/hour * 60 minutes/hour = $720/hour for babysitting. So she earned 720/hour - 12/hour = 608/hour - 12/hour = 486/hour for babysitting. As she worked for 50 minutes, she earned 486/hour * 50 minutes = 2430.
(C) 10. Rationale: Weng earns 12/60 = 0.2 per minute. Working 50 minutes, she earned 0.2*50 = 10.
(D) 1. Rationale: In an hour, there are 60 minutes. Weng did babysitting for 50 minutes, so she just worked for 50 / 60 = 1/12 of an hour. So she earned $12 * 1/12 = $1.
Correct Option: C

MCGSM8K-2Options
Question: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Options:
(A) 600. Rationale: Weng earns $12*50 = $600 for 50 minutes of babysitting.
(B) 10. Rationale: Weng earns 12/60 = 0.2 per minute. Working 50 minutes, she earned 0.2*50 = 10.
Correct Option: B

MCGSM8K-No-Rationale
Question: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Options:
(A) 600 (B) 2430 (C) 10 (D) 1
Correct Option: C

GSM8K-Judgement
Statement: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn? 10. Rationale: Weng earns 12/60 = 0.2 per minute. Working 50 minutes, she earned 0.2*50 =10.
Answer: True

Figure 2: Four example problems constructed from the original problem in GSM8K.

construct SFT data. RFT (Yuan et al., 2023) improves mathematical reasoning performance by applying CoT prompts and Rejection Sampling (RS) on SFT models to construct augmented solutions. MAmmoTH (Yue et al., 2023) utilizes a unique hybrid of CoT and program-of-thought (PoT) (Chen et al., 2022a) rationales to construct augmented solutions for improving mathematical problem-solving ability. As a result, these mathematical specialized models have surpassed previous general open-source LLMs by a significant margin in mathematical problem-solving.

Large Language Models for Mathematical Reasoning Process Evaluation. There are different ways to investigate the mathematical reasoning process, e.g., scoring each step of the reasoning process (Lightman et al., 2023), examining and analyzing each step, as well as modifying the reasoning process based on analyses (An et al., 2023). However, the above methods are mainly used to improve reasoning capabilities. Lightman et al. (2023) demonstrates that step-by-step verifying for the reasoning process of LLMs can lead to better model performance in mathematical problem-solving. An et al. (2023) fine-tunes LLMs on mistake-correction data pairs generated by GPT-4 to improve their reasoning capabilities in solving math problems. Existing research rarely explores the mathematical reasoning process evaluation ability of LLMs. We aim to design two types of benchmarks to test the ability of LLMs to evaluate the mathematical reasoning process of MWP. To achieve this goal, we design two direct and simple benchmarks, i.e., choosing the correct reasoning process from four candidates and judging whether a given reasoning process is correct or not. Please note that all existing mathematical reasoning benchmarks can be used to construct reasoning process evaluation benchmarks by our proposed method.

Similar to our work, there are also studies focusing on the counter-intuitive behaviors of LLMs. Berglund et al. (2023) studies the Reversal Curse of LLMs, i.e., LLMs exhibit a basic failure in deducting from “A is B” to the reverse direction “B is A”. Pezeshkpour and Hruschka (2023) aims to investigate the sensitivity of LLMs against options order of multiple-choice questions and recommends two approaches to calibrate LLMs' predictions including majority vote and multiple evidence calibration (MEC).

3 MCGSM8K

In this section, we describe in detail the process of constructing the benchmark MCGSM8K, which consists of multiple-choice questions with each question from the original test set of GSM8K and each option containing a solution. Our primary objective is to guarantee the reliability of MCGSM8K, ensuring that correct options are indeed correct and incorrect options are demonstrably wrong. To achieve this, we derive correct options directly from GSM8K, incorporating both the validated answer and its accompanying rationale, which are manually annotated to ensure accuracy. Conversely, incorrect options are generated by LLMs, comprising wrong answers distinct from the ground-truth answer, each accompanied by a plausible rationale. On the premise that the reliability of the dataset is guaranteed, we further enhance the challenge and difficulty of the dataset. Thus, we 1) use multiple advanced open-source LLMs to generate numerous candidate solutions with wrong answers, 2) maintain the diversity of incorrect option differences based on Rouge-L (Lin, 2005) scores and k-means cluster (Wong, 1979), and 3) ensure the confusability of the incorrect options with the correct option based on similarity ranking.

3.1 Generation of Incorrect Solutions

Given a question q_i, we utilize 8-shot CoT prompting to re-generate N solutions $\{(r_i^j, a_i^j) : j = 1, ..., N\}$ by advanced open-sourced models including Qwen (Bai et al., 2023), LLaMA-2 (Touvron et al., 2023b), MetaMath (Yu et al., 2023), and WizardLM (Xu et al., 2023), with sizes of 13B and 70B. We use each model to generate 50 candidate solutions. The principles for selecting these specific models are detailed in Appendix C. Specifically, a question q_i is appended to a few demonstrations, then fed to an LLM for generating its answer a_i^j along with rationale r_i^j step-by-step. In the generation process, we follow Wang et al. (2023b) to adopt temperature sampling and set the temperature as 0.7. Then, find out those with a wrong answer according to the ground-truth answer to construct the incorrect solution set $\mathrm{Set}_{\mathrm{aug}}$.

$$\begin{aligned}\mathrm{Set}_{\mathrm{aug}} =& \{(a_i^j, r_i^j) : \\ & a_i^j \neq a_i^*; i = 1, ..., K; j = 1, ...N\},\end{aligned} \tag{1}$$

where K is the total number of questions, and a_i^* is the ground-truth answer for question q_i.

3.2 Diversity and Quality

To enlarge diversity, an incorrect solution to q_i is added to $\mathrm{Set}_{\mathrm{aug}}$ only when its ROUGE-L similarity with any existing solutions to q_i in $\mathrm{Set}_{\mathrm{aug}}$ is less than 0.9, to effectively reconcile the retention quantity with the desired diversity of solutions. We also remove the calculation annotations proposed in the original answer rationale of GSM8K from each solution to facilitate formatting consistency. Invalid solutions are identified and filtered out based on heuristics, e.g., too-long or too-short solutions, solutions contain codes, and solutions do not end in the specified format. Finally, the candidate solutions to q_i are clustered into three categories using k-means clustering to ensure the diversity of solutions among each category. More details on data quality and diversity is provided in Appendix D.

3.3 Construction of Dataset

We have designed three different selection ways to avoid a single selection method causing a similar distribution or text of the three incorrect options. The ablation study on data construction is analyzed in Appendix B. One way is picking the one with the highest ROUGE-L similarity score to the correct option, which can ensure the confusability of the incorrect options. Another way is picking out the one with the lowest Perplexity (PPL) scored by language model WizardLM-70B (Xu et al., 2023). The lower the PPL, the solution is more natural and more consistent with the model's generation preferences. The last one is the random selection to keep diversity. For a question q_i, we select one solution from each of the three clusters to construct three incorrect options. At every election, we choose one way in turn. Then, the three incorrect options are combined with the correct one in a randomly shuffled order. Finally, we combine the question, options, and the correct option label into a multiple-choice question.

4 Experiments

In this section, the core of our analysis is how LLMs perform on the proposed benchmarks, what

Model	GSM8K	MCGSM8K	MCGSM8K -No-Rationale	MCGSM8K -2Options	GSM8K-Judgement TPR	TNR
			general open-source models			
LLaMA-2-13B	28.70	33.51	34.34	62.62	36.85	75.26
LLaMA-2-70B	56.80	38.29	58.45	66.94	47.08	78.09
AVG	42.75	35.90	46.40	64.78	41.97	76.68
			mathematical specialized models			
MetaMath-13B	72.30	23.12	45.49	57.70	85.75	17.54
MAmmoTH-13B	62.00	35.25	55.34	58.91	59.21	48.17
MetaMath-70B	82.30	34.87	76.04	64.22	88.93	22.21
MAmmoTH-70B	76.90	44.58	68.69	81.20	94.01	23.48
AVG	73.38	34.46	61.39	65.51	81.98	27.85
			general closed-source models			
GPT-3.5-Turbo	80.80	40.56	79.68	75.06	88.48	47.69
GPT-4	92.00	82.56	93.63	94.56	93.70	87.79
AVG	86.40	61.56	86.66	84.81	91.09	67.74
AVG All	68.98	41.59	63.96	70.15	74.25	50.03
RC Acc	0	25	25	50	50	50

Table 1: Comparison of testing accuracy on GSM8K and the proposed four benchmarks. To ensure equitable evaluations, we report the scores of all models using the settings of few-shot in-context learning. RC acc is the abbreviation of Random-Chance accuracy.

factors are associated with the model performance, and whether the performance can be further improved through supervised fine-tuning, thus answering the question: do LLMs master the ability to evaluate the mathematical reasoning process of MWP?

4.1 Can LLMs Solve Simple Mathematical Reasoning Process Evaluation of MWP?

To verify this issue, we propose a mathematical reasoning process evaluation benchmark MCGSM8K (Figure 2) consisting of 1,319 samples aiming at choosing the correct one from four solutions. We utilize the settings of few-shot in-context learning and CoT prompting, which is shown in **Appendix E** Figure 5.

4.1.1 Models

We evaluate the testing accuracy on several representative models – (i) general closed-source models GPT-4[2] (OpenAI, 2023) and GPT-3.5-Turbo (OpenAI, 2022), (ii) general open-source models: the current state-of-the-art LLaMA-2 (Touvron et al., 2023b) with two different parameter sizes of 13B and 70B, and (iii) mathematical specialized models: MetaMath (Yu et al., 2023) and MAmmoTH (Yue et al., 2023) in sizes 13B and 70B, which specifically tune LLaMA-2 on mathematical reasoning datasets. MetaMath is tuned on mathematical reasoning datasets collected by mathematical questions bootstrap and answering augmentation. MAmmoTH is trained on an instruction-tuning dataset compiled from 13 math datasets with a unique hybrid of CoT and PoT rationales. In all evaluation experiments, we set a temperature of zero for open-source models and mathematical specialized models following previous work (Yu et al., 2023; Yue et al., 2023), and a temperature of 0.2 for general closed-source models to generate quality answers.

4.1.2 Results

In Table 1, we cite the metrics of all tested models on GSM8K from the paper of MetaMath (Yu et al., 2023) and MAmmoTH (Yue et al., 2023). We can see that general closed-source models and mathematical specialized models have shown promising performance in GSM8K. For example, the accuracy of GPT-3.5-Turbo and MetaMath-70B all exceeds 80% from the third column in Table 1. These results exhibit the strong ability of most LLMs to

[2]We use `gpt4-1106-preview`.

solve grade school math word problems.

However, all models except GPT-4 achieve poor performance below 41% on the simple MCGSM8K compared with their performance on GSM8K. The accuracy of LLaMA-2-70B, MetaMath-70B, and GPT-3.5-Turbo drop from 56.80% to 38.29%, 82.30% to 34.87%, and 80.80% to 40.56%, respectively. To maximize accuracy, we test model accuracy with and without CoT prompting. As illustrated in the second column of Table 4 in **Appendix A**, the average performance gap between the test accuracy with and without CoT prompting is tiny, which is within 5%, exhibiting the incompetence of most LLMs in solving problems from MCGSM8K. We have tried different CoT prompting shown in Appendix Table 3. Specifically, we prompt the model to describe the task and explain the answer. The performance improvement brought by different prompts is less than 3%, and the best result so far is still much lower than that on GSM8K.

Point 1: Although most LLMs can solve MWP to some degree, they have difficulty in evaluating the reasoning process of MWP.

4.2 Can LLMs Solve MWP in the Form of Multiple-choice Questions?

In this subsection, we conduct the second experiment to investigate whether LLMs are incapable of solving multiple-choice format questions or evaluating the reasoning process. Specifically, for each option of MCGSM8K, we remove the reasoning process (rationale) and leave only the final answer to construct MCGSM8K-No-Rationale (Figure 2) consisting of 1319 samples. The model setting is the same as that in subsection 4.1.1. We report the testing accuracy under the settings of 5-shot in-context learning and CoT prompting, as shown in **Appendix E** Figure 5.

The average CoT performance of all tested models on MCGSM8K-No-Rationale is significantly higher than that on MCGSM8K (63.96% vs. 41.59%), and close to that on GSM8K (63.96% vs. 68.98%). From the paper of MAmmoTH (Yue et al., 2023), GPT-4 and MAmmoTH-70B achieve an accuracy of 72.60% and 65.00%, on the AQuA (Ling et al., 2017) dataset consisting of multiple-choice algebraic word problems, respectively. The results exhibit the ability of most LLMs to solve MWP in the form of multiple-choice questions. In addition, CoT prompting brings significant performance gains, e.g., the performance improvement of LLaMA-2-70B, and MetaMath-70B is 20.85%, and 38.74%, respectively, as shown in the third column of Table 4 in **Appendix A**.

Point 2: Most LLMs are capable of solving MWP in the form of multiple-choice questions.

4.3 Can Reducing Options on the Problem to Be Solved in MCGSM8K Improve Model Performance?

For each sample in MCGSM8K, we remove any two incorrect options from the four options, leaving only one correct and one incorrect option, resulting in MCGSM8K-2Options (Figure 2) consisting of 1319 samples. The model setup is the same as that in subsection 4.1.1. and we use 8-shot in-context learning without CoT prompting.

As illustrated in Table 1, the average accuracy of all tested models on the two-choice-question benchmark MCGSM8K-2Options is merely 1.17% higher than that on the open-formed-question benchmark GSM8K (70.15% vs. 68.98%), confirming that the ability of most LLMs to evaluate the reasoning process of MWP is insufficient.

Point 3: The ability of most LLMs to evaluate the reasoning process of MWP is insufficient.

4.4 Incapable of Identifying Correct Solutions or Incorrect Solutions?

From the previous experimental results, we observe that most LLMs perform poorly on MCGSM8K and MCGSM8K-2Options. To figure out whether the model is incapable of identifying correct solutions or incorrect solutions, we propose a true-or-false-question benchmark GSM8K-judgement (Figure 2) aiming at directly judging the correctness of a solution. Specifically, for each question in GSM8K, we append a solution to the end of the question. We utilize the correct solution synthesized by the ground-truth answer and the answer rationale for constructing the positive sample. For one open-formed question, there are theoretically infinite numbers of incorrect solutions generated by the model with each one varying from others. To eliminate randomness, we design three various negative samples utilizing the three high-quality incorrect solutions from the options of MCGSM8K, thus resulting in a total of 1319 positive samples and 3957 negative samples. The model setting is the same as that in subsection 4.1.1. We utilize 5-shot in-context learning and CoT prompts to maximize accuracy, as shown in **Appendix E** Figure 5.

Model	GSM8K	MCGSM8K	MCGSM8K -No-Rationale	MCGSM8K -2Options	GSM8K-Judgement TPR	TNR
LLaMA-2-13B	28.70	33.51	34.34	62.62	36.85	74.91
SFT-GSM8K	50.00	-	-	-	-	-
SFT-MCGSM8K	0.00	75.97	33.66	87.26	0.00	0.00
SFT-Judgement	0.00	22.21	27.82	51.60	69.60	83.24
SFT-hybrid	43.52	70.89	41.77	80.29	78.09	83.78

Table 2: Testing accuracy of LLaMA-2-13B trained on different data.

The testing accuracy of GSM8K-judgement is illustrated in the sixth and seventh columns of Table 1. As for general open-source models, the average True Negative Rate (TNR) is 76.68%, which is 34.71% higher than the average True Positive Rate (TPR) (41.97%) and 26.68% higher than the random-chance accuracy (50%). In general, it is easier for humans to identify incorrect solutions than correct solutions. This experimental result is consistent with human behavior. However, the performance of mathematical specialized models contradicts human performance, which are tuned on numerous augmented correct solutions. In contrast to general open-source models, the average TPR of mathematical specialized models is 40.01% higher, but the average TNR is 48.83% lower, which suggests that models merely fine-tuned on correct solutions can improve the ability to judge the correct reasoning process but greatly weaken the ability to judge the incorrect reasoning process. GPT-3.5-Turbo also shows similar performance as mathematical specialized models. Only GPT-4 achieves remarkable TPR and TNR results at the same time.

Point 4: Fine-tuning with only correct solutions improves the performance in evaluating correct solutions, but leads to a huge drop in evaluating incorrect solutions.

4.5 Fine-tuning

Finally, we explore whether fine-tuning can improve the ability of LLMs to evaluate the reasoning process of MWP. We use the most widely used model LLaMA-2-13B as the base model for fine-tuning. The LLaMA-2-13B is trained by fully fine-tuning on 8 NVIDIA A100 GPUs. The training details are following Yu et al. (2023).

In Table 2, the accuracy of SFT-GSM8K by fine-tuning the LLaMA-2-13B on the training data of GSM8K is extracted from RFT (Yuan et al., 2023). First, we fine-tune the base model on 6,000 training data from MCGSM8K and 3,000 training data from GSM8K-Judgement (1,500 positive samples and 1,500 negative samples), respectively. Significant performance improvements are obtained for both SFT-MCGSM8K and SFT-Judgement on in-domain datasets (IND), demonstrating the effectiveness of the training data in improving mathematical reasoning process evaluation ability. Meanwhile, SFT-Judgement achieves an improvement of 32.75% on TPR and 8.33% on TNR, revealing that finetuning on training data from GSM8K-Judgement is effective for enhancing capabilities in identifying both correct and incorrect solutions. On out-of-domain datasets (OOD) including GSM8K, MCGSM8K -No-Rationale, and GSM8K-Judgement, the performance decline between SFT-MCGSM8K and the base model can even reach up to 100%, which is consistent with the conclusion drawn from MAmmoTH (Yue et al., 2023) that fine-tuning LLMs using supervised data specific to certain datasets improves in-domain performance while reduces generalization to tasks beyond their fine-tuning data. Especially in the GSM8K benchmark, SFT-MCGSM8K and SFT-Judgement have completely lost their ability to follow instructions, resulting in irrelevant answers.

To maintain the generalization ability of the model in solving mathematical problems, we collect a hybrid fine-tuning dataset by mixing the training data from GSM8K, MCGSM8K, and GSM8K-Judgement. On GSM8K, SFT-hybrid lags behind SFT-GSM8K. We speculate that the question form of MCGSM8K and GSM8K-Judgement is completely different from that of GSM8K, resulting in the ability to evaluate the mathematical reasoning process not being successfully transferred to solve the mathematical reasoning problems. Compared with the base model, the improvement of the SFT-hybrid in TPR, TNR, MCGSM8K, and GSM8K is 41.24%, 8.87%, 37.38%, and 16.62%, respectively. Moreover, for the unseen MCGSM8K-No-Rationale task, the performance of the SFT-hybrid is better than the base model (+7.43%). The results show that we can improve the ability of LLMs to evaluate the mathematical reasoning process of MWP by fine-tuning.

4.6 Case Study

Appendix F shows some examples generated by different models to solve MCGSM8K and GSM8K-Judgement problems.

5 Conclusion

In this paper, we focus on exploring the ability of LLMs to evaluate the mathematical reasoning processes of MWP. To achieve this, we utilize incorrect solutions generated by multiple advanced LLMs to curate two benchmarks. One is MCGSM8K and its two variants, a new type of multiple-choice question dataset, in which each option contains a solution to solve the problem. The other one is GSM8K-Judgement, which judges whether a solution to a given problem is true or false. The poor performance of LLMs on MCGSM8K confirms the incapable ability of most LLMs in mathematical reasoning process evaluation. In particular, the performance on GSM8K-Judgement exhibits that it is easier to identify incorrect solutions. However, merely fine-tuning with correct solutions improves the performance in evaluating correct solutions, but leads to a huge drop in evaluating incorrect solutions. Fine-tuning models on the proposed training data greatly improves the mathematical process evaluation ability. Exploring the relation between the ability of mathematical problem-solving and mathematical reasoning process evaluation is left to future work.

Limitations

Through the above experiments and analyses, we summarize the following limitations:

1) In this work, we test the mathematical reasoning process evaluation ability of LLMs on limited benchmarks. In the future, we will utilize various MWP benchmarks, e.g., MATH and AQuA, to construct more comprehensive mathematical reasoning process evaluation benchmarks.

2) All incorrect solutions in the proposed benchmarks are generated by advanced LLMs, thus there are inevitably biases inherent in model generations. Furthermore, the constructed benchmark does not reflect the ability of LLMs to evaluate incorrect solutions written by humans.

3) The fine-tuned model does not exhibit a significant improvement in mathematical problem-solving. How to transfer the ability in mathematical reasoning process evaluation to mathematical problem-solving will be a future work.

Ethics Statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant 2023YFF1204901, the National Natural Science Foundation of China under Grant NSFC-62076172, and the Key Research and Development Program of Sichuan Province under Grant 2023YFG0116.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes LLM better reasoner. *arXiv preprint arXiv:2310.20689*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming 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, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022b. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 719–730. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison 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*.

C. Y Lin. 2005. Rouge-recall-oriented understudy for gisting evaluation-version 1.5.5.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

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*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2791–2809. Association for Computational Linguistics.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Singh Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3505–3523. Association for Computational Linguistics.

OpenAI. 2022. GPT-3.5 technical report.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:abs/2112.11446*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 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, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023a. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

J. A. Hartiganm. A. Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A The Impact of Different Prompts on Model Performance

First, we compare testing accuracy with/without Chain-of-Thought (CoT) prompting. We utilize CoT prompting to make models solve math problems through step-by-step natural language descriptions, as shown in Figure 5. On the MCGSM8K benchmark, the average accuracy of all tested models with and without CoT prompting is 41.59% and 38.46%, respectively, exhibiting that CoT prompting brings slight improvements. By analyzing CoT answers illustrated in Figure 6 and Figure 7, we find that most models have difficulty in identifying incorrect computational processes and logical fallacies in reasoning steps, thus leading to incorrect results. In addition to fine-tuning, some well-designed CoT prompts can also bring a certain degree of performance improvement.

On the MCGSM8K-No-Rationale benchmark, the accuracy with CoT prompting is significantly higher than the accuracy without CoT prompting for all tested models, e.g., the accuracy of LLaMA-2-70B with and without CoT prompting is 58.45% and 37.60%, and the accuracy of MetaMath-70B with and without CoT is 76.04% and 37.30%.

On the GSM8K-Judgement benchmark, the performance gap with and without CoT prompting is negligible for general closed-source models. Meanwhile, CoT prompting brings significant performance improvements for LLaMA-2-70B. We can observe a huge performance decline between CoT prompting and no CoT prompting for mathematical specialized models, as they forget the instruction-following ability without CoT prompting, resulting in completely irrelevant answers.

Model	Ours	P1
LLaMA-2-13B	33.51	33.43
LLaMA-2-70B	38.29	38.51
MetaMath-13B	23.12	23.58
MAmmoTH-13B	35.25	32.68
MetaMath-70B	34.87	35.71
MAmmoTH-70B	44.58	42.30

Table 3: Comparison of testing accuracy by different CoT prompts.

Further, we compare the testing accuracy by different CoT prompts on MCGSM8K. We utilize two different prompts including describing the task and explaining the answer (P1), and explaining why other options are wrong and why the predicted option is correct before outputting the correct answer label (ours). We illustrate the results in Table 3. The performance improvement brought by different prompts is less than 3%, and the best result so far is still much lower than that on GSM8K. We suspect this is mainly due to the model's high confidence in the incorrect solutions.

B Ablation Study on Data Construction

To keep the incorrect options in MCGSM8K confusing and diverse, we have designed three effective ways to select negative examples from candidate solutions, including random selection, similarity ranking, and PPL ranking. To explore the role of these three ways on the testing accuracy of the benchmark, we study the following setups:

(1) PPL Ranking (P): On the test data constructed by only PPL Ranking selection, LLaMA-2-70B achieves an accuracy of 51.60%, which is 13.31% higher than that on MCGSM8K, as shown in Figure 3. We suggest that the single PPL ranking selection may cause the constructed negative examples with high textual similarity. Thus, the performance gain may not be the improved ability of the model to evaluate the reasoning process, but rather the model successfully selecting the correct solutions based on text similarity.

(2) Random Selection + PPL Ranking (R + P): The performance of LLaMA-2-70B has dropped slightly from 51.60% to 49.28%. This ablation reflects that multiple selection methods are more likely to produce diverse negative examples, which can lead to improved performance. However, the quality of the negative samples by random selection cannot be guaranteed, thus the performance improvement is relatively small.

(3) Random selection + PPL Ranking + Similarity Ranking (MCGSM8K): After mixing similarity ranking, the performance of LLaMA-2-70B has dropped the most, reaching a minimum accuracy of 38.29%. We suggest that incorrect solutions similar to the correct solution are most likely to confuse the model.

C Principles for selecting LLMs to generate solutions

The reasons for choosing open-source LLMs and the inclusion of LLMs of relatively smaller sizes (i.e., 13B) to synthesize incorrect solutions are as follows:

Firstly, we need to emphasize that the stronger the model's mathematical problem-solving ability, the lower the probability of using reject sampling to get the desired incorrect answers. For example, we

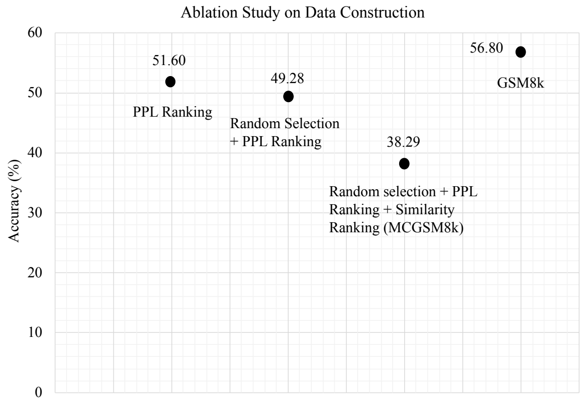

Figure 3: Testing accuracy of LLaMA-2-70B on the data constructed by different selecting ways.

found that because some GSM8K questions were too easy for the strong models such as GPT4, and MetaMath-70B, we couldn't get the desired wrong answer even to sample 50 times.

Secondly, we also tried to use prompt words to push chat models deliberately generating wrong answers, but we found some common patterns in the generated answers. Similarly, solutions generated from the same closed-source model also have similar characteristics and distributions, and these patterns are easily captured by LLMs, leading to a decline in the difficulty and quality of the benchmark.

Given the above factors, the efficiency of constructing incorrect solutions by closed-source models is low. To obtain a variety of diverse incorrect solutions for each question from GSM8K, we allow multiple LLMs with sizes of 13B and 70B including general LLMs Qwen, LLaMA-2, and WizardLM with stronger instruction following capabilities, and mathematical specialized LLMs MetaMath to generate incorrect solutions.

D MCGSM8K Analysis

D.1 Statistics

Table 5 describes the basic statistics of MCGSM8K, which consists of a total of 1,319 multiple-choice questions with each option containing one solution.

D.2 Diversity

We conduct further analysis to examine the distinctions among incorrect options. For each sample, we calculate the ROUGE-L overlap between the three incorrect options and illustrate the distribution of these scores in Figure 4. The results reveal significant diversity among incorrect options.

Model	MCGSM8K	MCGSM8K -No-Rationale	MCGSM8K -2Options	GSM8K-Judgement	
				TPR	TNR
LLaMA-2-13B	33.51/26.99	34.34/28.65	53.68/62.62	36.85/69.14	75.26/29.03
LLaMA-2-70B	38.29/33.66	58.45/37.60	65.13/66.94	47.08/10.99	78.09/66.72
AVG	35.90/30.33	46.40/33.13	59.41/64.78	41.97/40.07	76.68/47.88
MetaMath-13B	23.12/20.55	45.49/24.03	51.86/57.70	85.75/0.07	17.54/0.07
MAmmoTH-13B	35.25/26.16	55.34/27.37	54.89/58.91	59.21/16.22	48.17/16.15
MetaMath-70B	34.87/34.27	76.04/37.30	69.14/64.22	88.93/1.10	22.21/1.00
MAmmoTH-70B	44.58/40.86	68.69/35.25	69.07/81.20	94.01/1.70	23.48/5.80
AVG	34.46/30.46	61.39/30.99	61.24/65.51	81.98/4.77	27.85/5.76
GPT-3.5-Turbo	40.56/40.33	79.68/29.34	64.59/75.06	88.48/89.16	47.69/45.94
GPT-4	82.56/84.89	93.63/56.04	90.98/94.56	93.70/96.94	87.79/87.56
AVG	61.56/62.61	86.66/42.69	77.78/84.81	91.09/93.05	67.74/66.75
AVG All	41.59/38.46	63.96/34.45	64.92/70.15	74.25/35.67	50.03/31.53

Table 4: Comparison of testing accuracy with/without CoT prompting.

statistic	
# samples	1,319
avg. correct option length (in words)	54
avg. incorrect option length (in words)	55

Table 5: Statistics of the samples in the benchmark MCGSM8K.

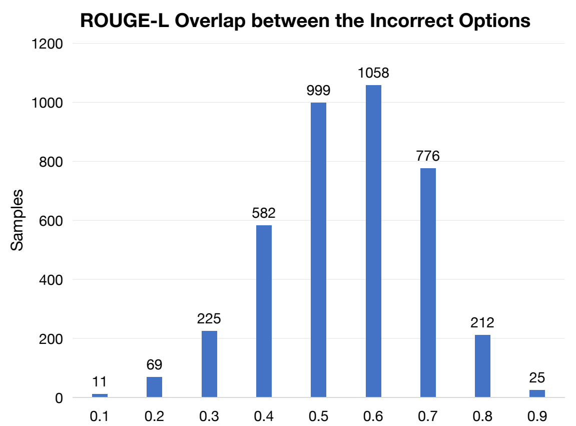

Figure 4: Distribution of the ROUGE-L scores between the incorrect options.

D.3 Quality

So far, we have demonstrated the quantity and diversity of incorrect options, yet their quality remains uncertain. To address this concern, we randomly select 100 incorrect options from MCGSM8K. Subsequently, three expert annotators (the authors of this work) are tasked with assigning a score ranging from 1 to 5 to each incorrect option, in terms of fluency, relevance, validity, and consistency. The evaluation results, presented in Table 6, indicate that the majority of incorrect options exhibit relevant and reasonable solutions to the given question, albeit some may contain a certain level of noise, such as repetition, calculation errors, etc. We use the intraclass correlation coefficient (CCI) to measure the correlation between raters. The correlation value calculated by ICC(C,1) is 0.92, higher than 0.9, which means that the evaluation of the three raters has a high degree of consistency.

score	1	2	3	4	5
distribution	1%	2%	15%	70%	12%

Table 6: Average human evaluation on the data quality for 100 incorrect options from MCGSM8K. 1: very poor, 2: poor, 3: okay, 4: good, and 5: very good.

E Prompts

We illustrate some examples in Figure 5.

F Case Study

F.1 MCGSM8K Case Study

We illustrate the result in Figure 6.

F.2 GSM8K-Judgement Case Study

We illustrate the result in Figure 7.

CoT Prompt with few-shot examples from MCGSM8K

The following are a math word problem and four solutions. Choose the correct one from the four options (A), (B), (C), and (D).

Question: Paddington has 40 more goats than Washington. If Washington has 140 goats, how many goats do they have in total?

Options: (A) 320. Rationale: If Washington has 140 goats, Paddington has 140+40 =180 goats. In total, they have 140+180 =320 goats (B) 240. Rationale: If Washington has 140 goats, and Washington has 40 more goats than Paddington, then Paddington has 140 - 40 =100 goats. Together, they have 100 + 140 =240 goats. (C) 400. Rationale: Washington has 140 goats + 40 goats =180 goats. Since Washington has 180 goats and Paddington has 40 more goats than Washington, Paddington has a total of 180 goats + 40 goats =220 goats. Together, they have 220 goats + 180 goats =400 goats. (D) 180. Rationale: If Washington has 140 goats, Paddington has 140+40 =180 goats. They have 180 goats in total.

Answer: Let's think step by step. Option (B) incorrectly calculates the number of goats Paddington has by subtracting 40 from the number of goats Washington has, which should be adding 40 to the number of goats Washington has.

Option (C) incorrectly calculates the number of goats Washington has by adding 40 to 140. This would give Washington 140 + 40 = 180 goats. However, the problem states that Washington has 140 goats.

Option (D) correctly calculates the number of goats Paddington has (180), but it incorrectly calculates the total number of goats.

Option (A) correctly calculates the total number of goats by adding the number of goats Washington has (140) to the number of goats Paddington has (180).

Therefore, the correct option is (A)

...

Question: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for $15 each. He ends up buying 8 video games for $20 each and has $5 left. How many lego sets does he still have?

Options: (A) 16. Rationale: He sold 13 lego sets because 5 / 5 = 1. He spent $100 on video games because 8 x 20 = 160. He earned $130 from the lego sets because 13 x 15 = 195. He had $235 worth of toys because 195 + 5 = 200. He had 17 lego sets because 200 / 13 = 15.385. He has 16 left because 15 + 1 = 16. (B) 10. Rationale: He earned $200 selling legos because 13 x 15 = 195. He spent $160 on video games because 8 x 20 = 160. He had $45 left over because 195 - 160 = 35. He bought 3 sets because 45 / 15 = 3. He now has 10 sets because 13 - 3 = 10. (C) 5. Rationale: He earned 13 * $15 = $200 from selling lego sets. He spent $200 - $10 = $190 on video games. He still has 13 - 8 = 5 lego sets. (D) 2. Rationale: He spent $160 on video games because 8 x 20 = 160. He earned $165 from his sale because 160 + 5 = 165. He sold 11 lego sets because 11 x 15 = 165. He has 2 sets left because 13 - 11 = 2.

Answer: Let's think step by step.

CoT Prompt with few-shot examples from MCGSM8K-No-Rationale

The following are a math word problem and four options. Choose the correct one from the four options (A), (B), (C), and (D).

Question: Paddington has 40 more goats than Washington. If Washington has 140 goats, how many goats do they have in total?

Options: (A) 320 (B) 240 (C) 400 (D) 180

Answer: Let's think step by step. If Washington has 140 goats, Paddington has 140+40 =180 goats. In total, they have 140+180 =320 goats. Therefore, the correct option is (A)

...

Question: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for $15 each. He ends up buying 8 video games for $20 each and has $5 left. How many lego sets does he still have?

Options: (A) 16 (B) 10 (C) 5 (D) 2

Answer: Let's think step by step.

CoT Prompt with few-shot examples from GSM8K-Judgement

The following is a statement about a math word problem. Please determine whether the statement is true or false.

Statement: Paddington has 40 more goats than Washington. If Washington has 140 goats, how many goats do they have in total? 320. Rationale: If Washington has 140 goats, Paddington has 140+40 =180 goats. In total, they have 140+180 =320 goats.

Answer: Let's think step by step. In this statement, it correctly calculates the total number of goats by adding the number of goats Washington has (140) to the number of goats Paddington has (180). The reasoning process and the final answer are correct.

Therefore, the statement is True.

...

Statement: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for $15 each. He ends up buying 8 video games for $20 each and has $5 left. How many lego sets does he still have? 2. Rationale: He spent $160 on video games because 8 x 20 = 160. He earned $165 from his sale because 160 + 5 = 165. He sold 11 lego sets because 11 x 15 = 165. He has 2 sets left because 13 - 11 = 2.

Answer: Let's think step by step.

Figure 5: CoT Prompts with few-shot examples from MCGSM8K, MCGSM8K-No-Rationale, and GSM8K-Judgement

MCGSM8K Case Study

Question: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for $15 each. He ends up buying 8 video games for $20 each and has $5 left. How many lego sets does he still have?

Options: (A) 16. Rationale: He sold 13 lego sets because 5 / 5 = 1. He spent $100 on video games because 8 x 20 = 160. He earned $130 from the lego sets because 13 x 15 = 195. He had $235 worth of toys because 195 + 5 = 200. He had 17 lego sets because 200 / 13 = 15.385. He has 16 left because 15 + 1 = 16. (B) 10. Rationale: He earned $200 selling legos because 13 x 15 = 195. He spent $160 on video games because 8 x 20 = 160. He had $45 left over because 195 - 160 = 35. He bought 3 sets because 45 / 15 = 3. He now has 10 sets because 13 - 3 = 10. (C) 5. Rationale: He earned 13 * $15 = $200 from selling lego sets. He spent $200 - $10 = $190 on video games. He still has 13 - 8 = 5 lego sets. (D) 2. Rationale: He spent $160 on video games because 8 x 20 = 160. He earned $165 from his sale because 160 + 5 = 165. He sold 11 lego sets because 11 x 15 = 165. He has 2 sets left because 13 - 11 = 2.

Ground-truth answer: D

LLaMA-2-70B answer: Option (A) is incorrect. The rationale states that he sold 13 lego sets because 5 / 5 = 1. This is incorrect. Then, it states that he spent $100 on video games because 8 x 20 = 160. This is also incorrect. Finally, it states that he earned $130 from the lego sets because 13 x 15 = 195. This is also incorrect.
Option (B) is incorrect. The rationale states that he earned $200 selling legos because 13 x 15 = 195. This is incorrect. Then, it states that he spent $160 on video games because 8 x 20 = 160. This is also incorrect. Finally, it states that he had $45 left over because 195 - 160 = 35. This is also incorrect.
Option (C) is correct. The rationale states that he earned 13 * $15 = $200 from selling lego sets. This is correct. Then, it states that he spent $200 - $10 = $190 on video games. This is also correct. Finally, it states that he still has 13 - 8 = 5 lego sets. This is also correct.
Option (D) is incorrect. The rationale states that he spent $160 on video games because 8 x 20 = 160. This is incorrect. Then, it states that he earned $165 from his sale because 160 + 5 = 165. This is also incorrect. Finally, it states that he sold 11 lego sets because 11 x 15 = 165. This is also incorrect.
Therefore, the correct option is (C)

MetaMath-70B answer: Option (A) is incorrect because it states that John sold 13 lego sets, but he actually sold 11 lego sets.
Option (B) is incorrect because it states that John spent $160 on video games, but he actually spent $160 on video games.
Option (C) is incorrect because it states that John has 5 lego sets left, but he actually has 2 lego sets left.
Option (D) is incorrect because it states that John sold 11 lego sets, but he actually sold 11 lego sets.
Therefore, the correct option is (C)

MAmmoTH-70B answer: 13 lego sets were sold because 5 / 5 = 1. He spent $100 on video games because 8 x 20 = 160. He earned $130 from the lego sets because 13 x 15 = 195. He had $235 worth of toys because 195 + 5 = 200. He had 17 lego sets because 200 / 13 = 15.385. He has 16 left because 15 + 1 = 16.
Therefore, the correct option is (A)

GPT-3.5-Turbo answer: Option (A) correctly calculates the amount of money John earned from selling the lego sets and the amount he spent on video games. However, it incorrectly calculates the number of lego sets he has left. The calculation should be 200/15 = 13.33, which means he has 13 lego sets left, not 16.
Option (B) correctly calculates the amount of money John earned from selling the lego sets and the amount he spent on video games. It also correctly calculates the number of lego sets he has left. Therefore, this option is correct.
Option (C) incorrectly calculates the amount of money John earned from selling the lego sets. It also does not provide any information about the number of lego sets he has left.
Option (D) incorrectly calculates the amount of money John earned from selling the lego sets. It also incorrectly calculates the number of lego sets he has left.
Therefore, the correct option is (B)

GPT-4 answer: Option (A) incorrectly calculates the total amount John earned from selling the lego sets and the total number of lego sets he had.
Option (B) incorrectly calculates the amount John had left over after buying the video games and the number of lego sets he bought.
Option (C) incorrectly calculates the amount John spent on video games and the number of lego sets he still has.
Option (D) correctly calculates the amount John spent on video games and the amount he earned from selling the lego sets. It also correctly calculates the number of lego sets he sold and the number he still has.
Therefore, the correct option is (D)

Figure 6: A comparison case on different types of models solving MCGSM8K problems.

GSM8K-Judgement Case Study

Statement: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for $15 each. He ends up buying 8 video games for $20 each and has $5 left. How many lego sets does he still have? 2.
Rationale: He spent $160 on video games because 8 x 20 = 160. He earned $165 from his sale because 160 + 5 = 165. He sold 11 lego sets because 11 x 15 = 165. He has 2 sets left because 13 - 11 = 2.

Ground-truth answer: True

LLaMA-2-70B answer: In the statement, it incorrectly calculates the total amount of money John spent on video games.
The final answer and reasoning process are incorrect. Therefore, the statement is False.

MetaMath-70B answer: In the statement, it correctly calculates the total amount of money John earned from selling his lego sets and the total amount of money he spent on video games. However, it incorrectly calculates the number of lego sets John still has.
The final answer and reasoning process are incorrect. Therefore, the statement is False.

MAmmoTH-70B answer: The final answer and reasoning process are correct. Therefore, the statement is True.

GPT-3.5-Turbo answer: In the statement, it correctly calculates the amount John spent on video games ($160) and the amount he earned from selling his lego sets ($165). However, it incorrectly calculates the number of lego sets he sold (11) and therefore the number of lego sets he still has.
The final answer and reasoning process are incorrect. Therefore, the statement is False.

GPT-4 answer: In the statement, it correctly calculates the total amount John spent on video games, the total amount he earned from selling his lego sets, and the number of lego sets he sold. It also correctly calculates the number of lego sets he has left.
The final answer and reasoning process are correct. Therefore, the statement is True.

Statement: Billy sells DVDs. He has 8 customers on Tuesday. His first 3 customers buy one DVD each. His next 2 customers buy 2 DVDs each. His last 3 customers don't buy any DVDs. How many DVDs did Billy sell on Tuesday?\n5. Rationale: Billy sells 3 DVDs to his first 3 customers. He sells 2 DVDs to his next 2 customers. He sells 0 DVDs to his last 3 customers. Billy sells a total of 3 + 2 + 0 = 5 DVDs on Tuesday..

Ground-truth answer: False

LLaMA-2-70B answer: In the statement, it correctly calculates the number of DVDs Billy sold to his first 3 customers (3 DVDs), the number of DVDs Billy sold to his next 2 customers (2 DVDs), and the number of DVDs Billy sold to his last 3 customers (0 DVDs).
The final answer and reasoning process are correct. Therefore, the statement is True.

MetaMath-70B answer: It correctly calculates the number of DVDs Billy sells to each group of customers and adds them up to find the total number of DVDs sold on Tuesday (5 DVDs).
The final answer and reasoning process are correct. Therefore, the statement is True.

MAmmoTH-70B answer: 3 + 2 = 5. Therefore, the statement is True.
5.

GPT-3.5-Turbo answer: In the statement, it correctly calculates the number of DVDs Billy sells to each group of customers and then adds them together to find the total number of DVDs sold (3 + 2 + 0 = 5).
The final answer and reasoning process are correct. Therefore, the statement is True.

GPT-4 answer: In the statement, it incorrectly calculates the total number of DVDs Billy sold on Tuesday. The first 3 customers bought 1 DVD each, making a total of 3 DVDs. The next 2 customers bought 2 DVDs each, making a total of 4 DVDs. The last 3 customers didn't buy any DVDs. Therefore, Billy sold a total of 3 + 4 = 7 DVDs on Tuesday.
The final answer and reasoning process are incorrect. Therefore, the statement is False.

Figure 7: A comparison case on different types of models solving GSM8K-Judgement problems.