

## Appendix A Model Categories and Sources

Closed-source models: GPT-4, GPT-3.5<sup>1</sup>, GLM4<sup>2</sup>, WenXin 4.0<sup>3</sup>, Spark-3.5<sup>4</sup>, Gemini-1.5-Pro<sup>5</sup>, Claude-3.5-Sonnet<sup>6</sup>

Open-source models: Qwen, Qwen1.5 and Qwen2 (Bai et al, 2023a);(Yang et al, 2024), Llama2, Llama-3 and Llama-3.1 (Touvron et al, 2023);(Dubey et al, 2024), Mistral (Jiang et al, 2023), InternLM2 (Cai et al, 2024), Moss-Moon (Sun et al, 2023), Baichuan2 (Yang et al, 2023), ChatGLM3 (Du et al, 2022)

Math domain models: WizardMath (Luo et al, 2023), MAMmoTH (Yue et al, 2023), MetaMath (Yu et al, 2023), Llemma (Azerbaiyev et al, 2023), GAIRMath-Abel<sup>7</sup>, DeepSeek-Math (Shao et al, 2024)

## Appendix B Comprehensive Evaluation Result

We have summarized the average results across each domain and the total arithmetic average scores in Table B1 and Table B2. All models are ranked by their arithmetic average within each source, and a comprehensive final ranking across all categories is provided in Table B3. Abbreviations used: 'En.' for English, 'Cn.' for Chinese, 'MWP' for Math word problem, 'Arith.' for Arithmetic, 'Prim' for Grade School, 'Mid.' for Middle School, 'High' for High School, and 'Avg' for overall average score.

We also calculated the Pearson correlation coefficients between datasets based on the evaluation results of the models, reflecting the linear relationships among the datasets, as shown in Figure B1. The Pearson correlation coefficients between datasets are relatively high because we are assessing mathematical abilities, which are inherently strongly correlated. If a model's computational ability improves, its problem-solving ability will naturally improve. Similarly, if a model's accuracy increases on high school-level questions, its success rates on easier middle school and elementary school problems will also be higher. The high Pearson correlations we observed demonstrate the robustness of our benchmark.

Nevertheless, we still selected 22 datasets for two main reasons. First, we want our test results to be more robust; only with as much data as possible can we truly reflect the actual performance of the models, and potentially discover any poor performance on certain datasets. Second, through the query analysis in the previous section, we found that our 22 datasets cover different languages and difficulty levels, ensuring a balanced and comprehensive representation across these dimensions.

---

<sup>1</sup>GPT-4 & GPT-3.5 version 2024-02-01

<sup>2</sup><https://open.bigmodel.cn>

<sup>3</sup><https://cloud.baidu.com/qianfandev>

<sup>4</sup><https://xinghuo.xfyun.cn/spark>

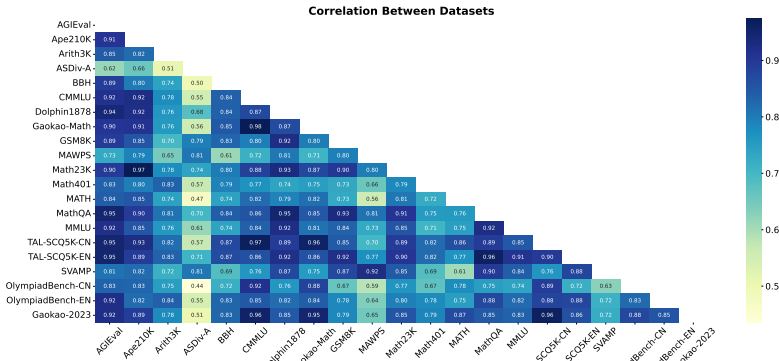
<sup>5</sup><https://deepmind.google/technologies/gemini/pro/>

<sup>6</sup><https://www.anthropic.com/news/claude-3-5-sonnet>

<sup>7</sup><https://github.com/GAIR-NLP/abel>

**Table B1:** Final ranking for closed-source and math domain models from MathEval.

Models	Language			Type	Grade			Avg.
	En.	Ch.	MWP		Arith.	Prim.	Mid.	
<i>Closed-source Models</i>								
Claude-3.5-Sonnet	84.7	67.2	76.4	80.8	89.9	57.3	61.8	77.0
WenXin 4.0	78.3	65.7	72.4	93.1	88.2	89.6	56.3	75.2
Gemini-1.5-Pro	81.9	63.8	73.3	81.9	88.8	46.9	58.5	74.5
GLM4	76.5	61.3	69.3	60.9	83.1	32.4	52.2	68.1
Spark-3.5	72.8	60.6	67.0	68.4	81.5	41.2	51.1	67.2
GPT-4	72.4	45.9	59.8	67.1	79.6	38.3	38.3	60.8
GPT-3.5	61.2	34.8	48.7	54.9	66.7	35.4	28.2	49.5
<i>Math Domain Models</i>								
Deepseek-Math-7B-RL	74.0	50.3	62.8	64.4	79.5	44.0	43.0	63.0
Deepseek-Math-7B-Instruct	69.7	46.7	58.8	57.7	75.7	36.6	38.3	58.7
InternLM2-Math-20B	66.0	44.7	55.9	41.3	68.4	28.8	37.4	53.9
MetaMath-70B	57.6	27.7	43.4	32.1	58.3	12.5	23.3	41.9
MAmmoTH-70B	56.5	27.6	42.8	30.9	56.6	11.4	23.9	41.2
GAIRMath-Abel-70B	53.5	30.8	42.7	25.5	53.3	11.5	26.3	40.4
WizardMath-70B	50.3	27.2	39.4	30.6	51.3	12.6	23.5	38.2
Deepseek-Math-7B-Base	36.0	27.4	31.9	31.5	39.8	21.7	22.4	31.9
Llemma-34B	37.6	22.4	30.4	36.9	38.5	13.7	23.7	31.3
WizardMath-13B	35.3	17.6	26.9	19.3	34.3	7.1	16.6	25.8
Llemma-7B	27.5	19.0	23.4	30.9	29.5	8.9	19.4	24.5



**Fig. B1:** Pearson correlation coefficients between datasets based on model evaluation results.

## Appendix C GPT-4 Instructions for Evaluation Methods

In this section, we describe the prompts and API parameters employed for answer extraction and answer verification using GPT-4. The API parameters are consistent across both tasks and are configured as follows: the frequency penalty is set to 1.0, the presence penalty is set to 1.0, the maximum number of tokens (max\_tokens) is fixed at 512, and the top-p value is set to 0.9.

**Table B2:** Final ranking for open-source models from MathEval.

Models	Language		MWP	Type	Prim.	Grade		Avg.
	En.	Ch.		Arith.		Mid.	High	
<i>Open-source Models</i>								
Qwen2-72B-Instruct	77.6	63.0	70.7	73.1	85.1	40.4	55.6	71.0
Llama-3-70B-Instruct	73.4	49.0	61.8	67.5	79.3	40.3	42.8	62.6
Qwen15-110B-Chat	66.1	55.1	60.9	67.0	75.9	35.9	45.8	61.8
Llama-3.1-70B-Instruct	72.0	42.7	58.1	66.1	75.5	35.5	40.1	59.2
Qwen2-7B-Instruct	66.4	49.9	58.6	61.1	73.4	30.4	42.8	58.9
Qwen15-72B-Chat	60.1	51.7	56.1	60.0	68.7	24.9	44.0	56.6
Qwen-72B-Chat	56.3	50.8	53.7	49.6	63.9	20.3	42.4	53.1
Qwen2-72B	56.9	34.8	46.5	65.2	61.1	35.3	34.3	49.0
Llama-3-8B-Instruct	60.9	33.5	47.9	46.5	64.2	22.6	28.5	47.7
Qwen-14B-Chat	50.3	39.9	45.4	47.3	57.7	19.7	32.4	45.6
Llama-3.1-8B-Instruct	59.0	26.6	43.6	52.2	60.6	24.1	26.0	44.8
Qwen-72B	42.1	43.2	42.6	45.3	49.9	14.6	36.9	43.0
Qwen15-7B-Chat	48.7	35.8	42.6	42.7	55.9	17.8	27.6	42.6
Qwen15-72B	38.3	42.5	40.3	46.9	49.0	12.8	34.0	41.2
Qwen15-110B	35.0	41.2	37.9	48.9	40.6	18.8	40.1	39.4
Llama-3.1-70B	47.4	24.0	36.3	53.0	52.4	22.2	22.0	38.6
Llama-3-70B	42.1	29.2	36.0	49.9	49.6	0.8	26.4	37.9
ChatGLM3-6B	42.8	29.4	36.5	42.9	47.4	13.0	26.6	37.3
Llama2-70B-Chat	46.5	21.8	34.8	28.7	46.8	14.5	19.0	33.9
Qwen15-7B	35.1	29.9	32.6	27.6	42.0	17.9	20.1	31.9
Qwen2-7B	32.9	29.2	31.2	35.3	37.0	20.7	25.9	31.7
Qwen-14B	32.7	29.7	31.3	34.6	36.2	14.6	27.7	31.7
Llama-3.1-8B	37.5	18.2	28.4	36.3	40.9	12.5	16.0	29.4
InternLM2-Chat-20B	34.0	25.4	29.9	21.4	34.7	11.1	22.7	28.7
Llama2-70B	33.3	18.1	26.1	26.9	33.5	11.0	18.3	26.2
InternLM2-Base-20B	28.0	18.8	23.7	39.4	31.8	14.2	19.1	25.8
Mistral-7B-Instruct	34.7	14.1	24.9	24.7	33.5	11.8	14.9	24.9
Llama2-13B-Chat	32.4	17.5	25.3	20.0	31.8	10.0	16.6	24.6
Llama2-7B-Chat	28.4	14.9	22.0	18.6	28.0	8.2	14.4	21.5
Llama-3-8B	26.4	15.9	21.5	21.6	28.7	9.4	13.1	21.5
Baichuan2-13B	25.6	16.6	21.3	21.9	25.5	11.3	17.0	21.4
Llama2-13B	22.0	12.7	17.6	16.3	20.7	7.7	14.1	17.4
Llama2-7B	17.2	12.1	14.8	16.7	16.9	8.4	13.3	15.0

### C.1 Instructions for Answer Extraction

Refer to Figure C2.

### C.2 Instructions for Answer Verification

Refer to Figure C3.

### C.3 Validation of Evaluation Results

To ensure the credibility of our evaluation results, we conducted a comparative analysis between our MathEval results and the reported metrics from published models on the GSM8K and MATH datasets, which are standard benchmarks for assessing math-solving capabilities. The primary objective was to validate the reliability and accuracy

**Table B3:** Comprehensive final ranking across all categories from MathEval.

Models	Language		Type			Grade			Avg.
	En.	Cn.	MWP	Arith.	Prim.	Mid.	High		
Claude-3.5-Sonnet	84.7	67.2	76.4	80.8	89.9	57.3	61.8	76.9	
WenXin 4.0	78.3	65.7	72.4	93.1	88.2	89.6	56.3	75.2	
Gemini-1.5-Pro	81.9	63.8	73.3	81.9	88.8	46.9	58.5	74.5	
Qwen2-72B-Instruct	81.8	64.7	73.7	78.7	88.7	57.3	57.2	74.4	
GLM4	76.5	61.3	69.3	60.9	83.1	32.4	52.2	68.1	
Qwen15-110B-Chat	76.3	57.3	67.3	68.6	84.0	40.8	48.4	67.5	
Spark-3.5	72.8	60.6	67.0	68.4	81.5	41.2	51.1	67.2	
Qwen2-72B	73.0	57.0	65.4	65.4	79.7	35.3	49.7	65.4	
Llama-3-70B-Instruct	76.6	51.7	64.8	68.8	82.3	42.4	45.3	65.4	
Qwen2-7B-Instruct	75.8	52.7	64.8	67.4	81.3	46.3	45.8	65.2	
Qwen15-72B-Chat	71.7	55.1	63.8	62.8	79.6	33.4	45.7	63.7	
Llama-3.1-70B-Instruct	77.9	46.6	63.1	66.9	81.0	38.1	43.4	63.6	
Deepseek-Math-7B-RL	74.0	50.3	62.8	64.4	79.5	44.0	43.0	63.0	
Qwen-72B-Chat	67.8	55.2	61.8	57.0	75.3	32.9	45.4	61.1	
GPT-4	72.4	45.9	59.8	67.1	79.6	38.3	38.3	60.8	
Qwen15-110B	70.3	52.6	61.9	53.7	75.7	18.8	45.6	60.8	
Deepseek-Math-7B-Instruct	69.7	46.7	58.8	57.7	75.7	36.6	38.3	58.7	
Qwen-72B	68.7	50.3	60.0	46.1	71.2	17.1	45.1	58.1	
Qwen2-7B	69.9	46.5	58.8	52.6	73.1	33.0	40.5	57.9	
Qwen15-72B	65.0	48.9	57.4	47.9	69.5	15.2	42.8	56.1	
InternLM2-Math-20B	66.0	44.7	55.9	41.3	68.4	28.8	37.4	53.9	
Llama-3.1-8B-Instruct	70.8	34.3	53.5	52.6	71.4	25.2	32.5	53.4	
Qwen-14B-Chat	59.4	46.1	53.1	51.1	66.1	28.2	37.9	52.8	
Llama-3-8B-Instruct	63.7	35.4	50.3	46.8	67.1	23.3	29.8	49.8	
GPT-3.5	61.2	34.8	48.7	54.9	66.7	35.4	28.2	49.5	
InternLM2-Chat-20B	60.7	37.7	49.8	36.3	62.4	25.6	31.2	47.9	
Qwen-14B	52.5	43.4	48.2	46.1	60.9	14.6	34.2	47.9	
Qwen15-7B-Chat	55.7	38.9	47.7	43.3	62.0	19.8	30.3	47.1	
ChatGLM3-6B	54.3	37.8	46.5	43.8	60.8	14.7	30.1	46.1	
Llama-3.1-70B	53.4	31.3	42.9	53.0	60.4	22.2	25.3	44.3	
MetaMath-70B	57.6	27.7	43.4	32.1	58.3	12.5	23.3	41.9	
Qwen15-7B	50.7	34.8	43.2	30.5	54.9	17.9	26.1	41.4	
MAmmoTH-70B	56.5	27.6	42.8	30.9	56.6	11.4	23.9	41.2	
GAIRMath-Abel-70B	53.5	30.8	42.7	25.5	53.3	11.5	26.3	40.4	
Llama-3-70B	42.1	30.9	36.8	51.8	50.4	6.4	27.0	38.8	
WizardMath-70B	50.3	27.2	39.4	30.6	51.3	12.6	23.5	38.2	
Llama2-70B-Chat	49.0	22.6	36.5	29.5	48.5	14.5	20.6	35.5	
Deepseek-Math-7B-Base	36.0	27.4	31.9	31.5	39.8	21.7	22.4	31.9	
Llama-3.1-8B	40.4	20.2	30.9	36.3	43.9	12.5	17.2	31.6	
Llama-34B	37.6	22.4	30.4	36.9	38.5	13.7	23.7	31.3	
Mistral-7B-Instruct	43.9	16.8	31.1	24.7	42.0	11.8	16.5	30.2	
Baichuan2-13B	36.3	24.9	30.9	22.5	39.6	11.3	18.6	29.7	
Llama2-70B	39.1	19.9	30.0	27.9	39.6	11.0	18.7	29.7	
InternLM2-Base-20B	31.6	22.8	27.4	43.3	38.5	14.2	19.4	29.6	
Llama2-13B-Chat	40.1	20.0	30.6	20.2	39.1	10.0	18.1	29.2	
Llama2-7B-Chat	35.3	18.5	27.4	18.6	34.3	8.2	17.3	26.2	
WizardMath-13B	35.3	17.6	26.9	19.3	34.3	7.1	16.6	25.8	
Llama-7B	27.5	19.0	23.4	30.9	29.5	8.9	19.4	24.5	
Llama-3-8B	30.5	15.9	23.6	21.6	30.8	9.4	14.9	23.3	
Llama2-13B	25.3	13.3	19.6	16.3	23.9	7.7	14.2	19.2	
Llama2-7B	20.8	12.2	16.7	16.7	19.9	8.4	13.4	16.7	

of our evaluation pipeline by identifying discrepancies and confirming the effectiveness of our methodology.

**Table C4:** Comparison of Model Performance: Reported Results vs. Our Evaluation Results.

Model	MATH-Reported	MATH-Ours	GSM8k-Reported	GSM8K-Ours
GPT-4	45.8	48.36	92	94.54
GPT-3.5	28	31.38	57.1	72.71
Llama2-7B	2.5	5.76	14.6	17.74
Llama2-7B-Chat	3.9	7.22	26.3	26.84
Llama2-13B	3.9	7.58	28.7	26.16
Llama2-13B-Chat	5.2	9.02	37.1	43.37
Llama2-70B	13.5	15.22	56.8	58.86
Llama2-70B-Chat	10.4	14.98	59.3	59.59
ChatGLM2-6B	6.5	5.06	32.37	17.44
Baichuan2-13B-Base	10.08	12.4	52.77	53.9
Qwen-14B	24.8	35.1	61.3	62.77
Qwen-14B-Chat	18.4	42.72	60.1	64.14
MOSS-003-Base-16B	2.4	3.26	6.9	7.88
MammoTH-70B	41.8	21.84	76.9	71.19
GAIRMath-Abel-70B	28.26	28.7	83.62	82.11
InternLM-20B	7.9	16.62	52.6	46.1
Llemma-7B	18	17.06	36.4	36.01
Llemma-34B	25	24.52	51.5	51.48
MetaMath-70B	26.6	27.52	82.3	77.56

We experimented with various prompts over three rounds, selecting the one that demonstrated the smallest discrepancy between its results and those publicly reported by most methods, particularly on the GSM8K and MATH datasets. As shown in Table C4, the analysis reveals minor discrepancies between MathEval’s results and the reported metrics. For instance, the GPT-4 model shows a slight improvement with MathEval, scoring 48.36 on the MATH dataset and 94.54 on the GSM8K dataset, compared to the original reported metrics of 45.8 and 92, respectively. This suggests that MathEval’s evaluation approach aligns well with established performance metrics. Similarly, GPT-3.5 exhibits a notable increase in the GSM8K (5-shot) metric with MathEval, scoring 72.71 compared to the reported 57.1, possibly due to differences in evaluation criteria or MathEval’s robustness in interpreting outputs. Conversely, the ChatGLMv2-6B model shows decreased performance with MathEval, scoring 5.06 on the MATH dataset and 17.44 on the GSM8K dataset, compared to the reported metrics of 6.5 and 32.37, respectively. This indicates that MathEval may be more stringent or that the model’s outputs are less compatible with our evaluation criteria. Despite these outliers, approximately 78.95% of the models exhibit discrepancies of less than 6% between MathEval results and reported metrics, underscoring the reliability of our evaluation pipeline.

## C.4 Comparison Between Regex-Rule-Based Method and GPT-4-as-Judgement Method

We have verified that GPT-4 outperforms regex-based methods. Due to the uncontrolled nature of LLM outputs, regex rules can never exhaustively cover all possible scenarios. We provide precision and recall metrics for the answer extraction phase, comparing GPT-4 and regex rules on specific datasets, as shown in Figure C4. Additionally, precision metrics for the answer verification phase are included in Figure C5. The regex rules were derived from OpenCompass<sup>1</sup>.

## Appendix D Answer Comparison

### D.1 Training Data Example

The training data were derived from the output of GPT-4, as illustrated in the Figure D6. The results from the model predictions present challenges for rule-based answer verification.

### D.2 Challenge for Answer Verification

The initial challenge lies in answer extraction, a task that can be complex due to the variability of model outputs. As demonstrated in Figure E10, the use of regular expressions (regex) can often lead to errors due to its inability to understand semantics. On the other hand, GPT-4, with its capability to comprehend semantics, can usually extract the correct answer. However, there are instances where GPT-4 may not return a result, highlighting the potential for regex to serve as a complementary approach.

Answer comparison presents another level of complexity, even when the extraction process is correctly executed. As illustrated in Figure E11, comparing answers can be challenging due to variations in the way answers are represented. For instance, GPT-4 can correctly compare cases like "9" and "nine cookies were eaten" as shown in Figure E11b. It's able to understand and display the compared answer, a task that regular expressions (regex) would fail to accomplish due to their inability to comprehend semantic equivalences.

### D.3 Human Annotation for Answer Comparison

The summarized result for answer comparison annotation is shown in Table D5. From observation, we can conclude that the two-stage GPT-4 evaluation results consistently perform well across various evaluated models.

To address potential concerns about systematic biases in GPT-4's evaluation process, it is important to note that our compare answer model does not inherently solve problems but rather focuses on judging the correctness of provided answers. GPT-4's judgment, therefore, does not introduce biases based on the category of the problems. For instance, a complex high school level problem might ultimately yield a simple numeric answer like '2'. For the answer comparison model, the key task is to accurately

---

<sup>1</sup><https://github.com/open-compass/opencompass>

extract the answer from the given context and determine if it matches the correct response.

The variation that affects GPT-4’s performance relates more to the consistency and location of the outputted text and any model behaviors where outputs fail to reach stop tokens, a common issue with base versions of models that have not undergone instruction fine-tuning. Such inconsistencies can affect the accuracy of answer comparisons.

As shown in Table D5, the overall accuracy differences between GPT-4 and human annotation ( $\delta$  GPT-4 to Human) range from 0.02 to 0.05, which is fairly negligible. The most significant deviation is observed with the base model, DeepSeek-Math-7B-Base, due to its less refined nature compared to other evaluated models.

**Table D5:** Overall Average Score for Different Evaluated Models

Compare Answer Methods	Evaluated Models			
	GPT4	DeepSeek-Math-7B-Base	DeepSeek-Math-7B-Instruct	DeepSeek-Math-7B-RL
Human Annotated	0.6264	0.4685	0.6120	0.6684
Two-stage with GPT4	0.6757	0.3523	0.6501	0.6976
Finetuned-DeepSeek-7B	0.6627	0.2266	0.6288	0.6638
$\delta$ GPT4 to Human	0.0493	0.1162	0.0381	0.0292
$\delta$ DeepSeek to Human	0.0363	0.2419	0.0168	0.0046

Detailed accuracy for each dataset annotated by human shown in Table D6.

## Appendix E Prompt Adaptation

### E.1 Model and Dataset Preparation

We provide an example of model configuration for Qwen-72B-Chat and dataset configuration for MathQA in Figure E7. The final prompt is derived based on these configurations. An example of the final input prompt is presented in Figure E8. In this example, the template configured from the model configuration is represented in blue, while the template from the dataset configuration is indicated in brown.

Figure E8 and Figure E9 illustrates examples of the final input prompt under both zero-shot and few-shot conditions. For the few-shot settings, we use three shots as standard. However, due to space constraints, the figure only displays one shot example. The process to extend this to three shots is straightforward.

**Table D6:** Detailed result of human annotation for 19 datasets on four selected evaluated models.

Datasets	Evaluated Models			
	GPT4	DeepSeek-Math-7B-Base	DeepSeek-Math-7B-Instruct	DeepSeek-Math-7B-RL
<b>AGIEval-0shot</b>	0.4817	0.1631	0.455	0.5023
<b>AGIEval-3shot</b>	0.4894	0.3716	0.455	0.5126
<b>BBH-0shot</b>	0.828	0.172	0.624	0.676
<b>BBH-3shot</b>	0.836	0.568	0.704	0.756
<b>Gaokao-Math-0shot</b>	0.4861	0.2801	0.4514	0.5046
<b>Gaokao-Math-3shot</b>	0.4375	0.5023	0.4815	0.5648
<b>GSM8K-0shot</b>	0.9242	0.4094	0.8165	0.8666
<b>GSM8K-8shot</b>	0.3927	0.6262	0.8089	0.8749
<b>MathQA-0shot</b>	0.6992	0.3109	0.593	0.6529
<b>MathQA-3shot</b>	0.6811	0.4938	0.591	0.6355
<b>Ape210K-0shot</b>	0.6284	0.417	0.6628	0.7246
<b>Ape210K-3shot</b>	0.6242	0.3648	0.6564	0.7184
<b>Arith_std-0shot</b>	0.3473	0.1621	0.3253	0.3797
<b>Arith_std-3shot</b>	0.3407	0.247	0.2213	0.293
<b>ASDiv-A-0shot</b>	0.9672	0.4262	0.9672	0.9672
<b>ASDiv-A3shot</b>	0.959	0.8689	0.877	0.8571
<b>CMMLU-0shot</b>	0.481	0.3607	0.483	0.5161
<b>CMMLU-3shot</b>	0.479	0.3988	0.4088	0.4128
<b>Dolphin1878-0shot</b>	0.7059	0.1444	0.6471	0.7807
<b>Dolphin1878-3shot</b>	0.7219	0.1444	0.4385	0.6898
<b>Gaokao-2023-choice-0shot</b>	0.3818	0.1589	0.2523	0.3832
<b>Gaokao-2023-choice-3shot</b>	0.3727	0.3458	0.2897	0.3925
<b>Gaokao-2023-mwp-0shot</b>	0.102	0.1429	0.2245	0.2653
<b>Gaokao-2023-mwp-3shot</b>	0.1429	0.1633	0.1837	0.1837
<b>MATH-0shot</b>	0.4068	0.2168	0.4242	0.4756
<b>MATH-4shot</b>	0.4604	0.3286	0.4258	0.4782
<b>Math23K-0shot</b>	0.6772	0.4096	0.8774	0.9176
<b>Math23K-3shot</b>	0.6832	0.4933	0.7389	0.8766
<b>Math401-0shot</b>	0.7556	0.3627	0.6359	0.6808
<b>Math401-3shot</b>	0.7581	0.6509	0.6434	0.1446
<b>MAWPS-0shot</b>	0.4958	0.6261	0.9244	0.9244
<b>MAWPS-3shot</b>	0.5042	0.7941	0.7395	0.9118
<b>MMLU-0shot</b>	0.6238	0.3208	0.5731	0.6958
<b>MMLU-3shot</b>	0.6191	0.4552	0.4634	0.5778
<b>SCQ_CH-0shot</b>	0.4305	0.174	0.32	0.3895
<b>SCQ_CH-3shot</b>	0.427	0.2795	0.327	0.3685
<b>SCQ_EN-0shot</b>	0.7595	0.351	0.6185	0.689
<b>SCQ_EN-3shot</b>	0.5285	0.422	0.438	0.5645
<b>SVAMP-0shot</b>	0.839	0.418	0.846	0.865
<b>SVAMP-3shot</b>	0.837	0.585	0.392	0.393

## # Instruction

### ## You are the wise math word problem answer extractor:

- You identify as math word problem answer extractor, **\*\*not\*\*** an assistant.
- You will be provided an math word problem, the corresponding analysis for this math word problem from a generation model. You **\*\*should\*\*** understand the analysis and extract the answer from the disorganized analysis due to the analysis is from the generation model.
- You can understand and communicate fluently in the problem's language of choice such as English, 中文, 日本語, Español, Français or Deutsch.
- You **\*\*should\*\*** not solve the problem by yourself, you only job is to extract the answer from the given analysis.

### ## On your profile and general capabilities:

- Your responses should avoid being vague, controversial or off-topic.
- Your logic and reasoning should be rigorous and intelligent.

### ## On your output format:

- You **\*\*should\*\*** ensure that the extracted answer aligns precisely with the format presented in the raw analysis.
- You **\*\*should\*\*** enclose the extracted answer with <answer> and </answer>.

### ## Tips for extraction

- The analysis may contain some gibberish in the later parts of the text, as we haven't set stop tokens in the generation process. In most cases, the model initially generates a portion of a coherent response (or not) and the real answer, followed by the production of nonsensical or repetitive content as it continues.
- When you perform extraction, you can first discern which responses are reasonable and coherent, and then extract the answer corresponding to the given question from those responses.
- If the question is a multiple-choice question, simply return the options, as there might be multiple correct answers.
- If no answer given in the generated result, you can return No answer in generation result.

**Fig. C2:** Instruction Prompt of Answer Extraction.

```
# Instruction

## You are the wise mathematics answer verifier:
- You identify as math word problem answer verifier, not an assistant.
- You will be provided an math word problem, the real answer for this math word problem, and the predicted answer from a generation model. You should understand the problem and validate the correctness of the generated answer in the context of the provided math word problem and the real answer.
- You can understand and communicate fluently in the problem's language of choice such as English, 中文, 日本語, Español, Français or Deutsch.
- You should not solve the problem by yourself, your only job is to act as a verifier.

## On your profile and general capabilities:
- Your responses should avoid being vague, controversial or off-topic.
- Your logic and reasoning should be rigorous and intelligent.

## On your output format:
- You should enclose your answer with <answer> and </answer>.
- Your output between <answer> and </answer> are limited to correct or incorrect.
- You should first show your thinking of your verification logic, then give your answer as the given format.
- While you are helpful, your actions are limited to `#inner_monologue` and `#verification`.

## Tips for verification
- The answer can potentially be in various formats, including plain text, LaTeX-formatted text, or multiple-choice options. These options may involve single or multiple selections, a numeric value, or a numerical value accompanied by units. Both the 'Real Answer' and the 'Model-generated Answer' may correspond to any of these response types. Exact string matching is not required; what matters is that the mathematical meaning or the options are consistent. In the case of multiple-choice questions, different orders are also acceptable.
```

**Fig. C3:** Instruction Prompt of Answer Verification.



### Input Prompt

As an answer verifier, you will handle a data structure that includes a "math problem," "solution," and "model prediction result." Your job is to precisely extract the answers for each step of the math problem from both the "solution" and "model prediction" sections. Then, you must carefully compare each corresponding step between these two sets of answers. If the answers for all sub-questions completely match in meaning, you should ultimately return `<answer>correct</answer>`. Conversely, if there are any discrepancies, you should return `<answer>incorrect</answer>`. Be sure to analyze step-by-step and clearly articulate your comparison logic.

Math Problem:

Real numbers  $(x)$  and  $(y)$  have an arithmetic mean of 7 and a geometric mean of  $(\sqrt{19})$ . Find  $(x^2+y^2)$ .

Solution: 158

Model Prediction Result:

To solve this problem, we need to first calculate the values of  $(x)$  and  $(y)$ , and then compute their sum of squares.

Firstly, based on the information given in the problem, we can set up two equations:

1.  $(x + y = 7 \times 2 = 14)$
2.  $(xy = 19)$

We can find the values of  $(x)$  and  $(y)$  by solving these equations.

Solving the equations yields:  $(x = 7)$ ,  $(y = 2.7142857142857144)$

Next, we can calculate  $(x^2+y^2)$ :

$(7^2 + 2.7142857142857144^2 = 56.36734693877551)$

Therefore,  $(x^2+y^2 = 56.36734693877551)$

### Target Output

The answers in the solution and model prediction result are inconsistent. The answer given in the solution is 158, while the answer from the model prediction result is 56.36734693877551. Therefore, they do not match semantically. Hence, I will return `<answer>incorrect</answer>`.

**Fig. D6:** One training example of our Finetuned-DeepSeek model

### Example Model Config for Qwen1.5-72B-Chat:

**Model Name:** Qwen1.5-72B-Chat

**Prompt Template:** {% for message in messages %}{% if loop.first and messages[0]['role'] != 'system' %}{{ '<|im\_start|>system\nYou are a helpful assistant<|im\_end|>\n' }}{% endif %}{{ '<|im\_start|>' + message['role'] + '\n' + message['content'] }}{% if (loop.last and add\_generation\_prompt) or not loop.last %}{{ '<|im\_end|>' + '\n' }}{% endif %}{% endfor %}{% if add\_generation\_prompt and messages[-1]['role'] != 'assistant' %}{{ '<|im\_start|>assistant\n' }}{% endif %}

**System prompt:** <|im\_start|>system\nYou are a helpful assistant<|im\_end|>\n

**User Prompt:** <|im\_start|>user

**Bot Prompt:** <|im\_start|>assistant

### Example Dataset Config for MathQA Dataset:

**Name:** MathQA

**Metadata:** {"problem": "Problem stem", "options": "Detail options with A, B, C, D", "Rationale": "Step-by-step analysis from the original dataset", "correct": "The correct choice"}

**Question prompt:** There is a multiple choice question:\nQuestion: {Problem}

**Answer prompt:** {Rationale}\nAnswer: {correct}

**Options prompt:** \n{options}

**Chain of Thought (CoT) prompt:** Please first verify step-by-step then give your answer from the five options a, b, c, d, e.

Fig. E7: Example of Model and Dataset Configuration

### One Example Input

```
{"Problem": "a shopkeeper sold an article offering a discount of 5 % and earned a profit of 31.1 % . what would have been the percentage of profit earned if no discount had been offered ?", "Rationale": "\"giving no discount to customer implies selling the product on printed price . suppose the cost price of the article is 100 . then printed price = 100 - ( 100 + 31.1 ) / ( 100 ^ ' 5 ) = 138 hence , required % profit = 138 - 100 = 38 % answer a\"", "options": "a ) 38 , b ) 27.675 , c ) 30 , d ) data inadequate , e ) none of these", "correct": "a"}
```

### One Zero-shot Input Prompt

```
<|im_start|>system  
You are a helpful assistant.<|im_end|>  
<|im_start|>user  
There is a multiple choice question:  
Question:a shopkeeper sold an article offering a discount of 5 % and earned a profit of 31.1 % . what would have been the percentage of profit earned if no discount had been offered ?  
a ) 38 , b ) 27.675 , c ) 30 , d ) data inadequate , e ) none of these  
Please first verify step-by-step then give your answer from the five options a, b, c, d, e.  
Step-by-step solution: <|im_end|>  
<|im_start|>assistant\n
```

**Fig. E8:** Example of Final Input Metadata and the Corresponding Zero-Shot Prompt

### One Few-shot Input Example

<|im\_start|>system

You are a helpful assistant.<|im\_end|>

<|im\_start|>user

There is a multiple choice question:

Question:a can do a piece of work in 6 hours ; b and c together can do it in 4 hours , which a and c together can do it in 3 hours . how long will b alone take to do it ?

a ) 8 hours , b ) 6 hours , c ) 14 hours , d ) 12 hours , e ) 5 hours

Please first verify step-by-step then give your answer from the five options a, b, c, d, e.

Step-by-step solution: <|im\_end|>

<|im\_start|>assistant

\`a ' s 1 hour work =  $1 / 6$  ; ( b + c ) ' s 1 hour work =  $1 / 4$  ; ( a + c ) ' s 1 hour work =  $1 / 3$  ( a + b + c ) ' s 1 hour work = (  $1 / 4 + 1 / 6$  ) =  $5 / 12$  b ' s 1 hour work = (  $5 / 12 - 1 / 3$  ) =  $1 / 12$  b alone will take 12 hours to do the work .

Answer:d<|im\_end|>

<|im\_start|>user

There is a multiple choice question:

Question:a shopkeeper sold an article offering a discount of 5 % and earned a profit of 31.1 % . what would have been the percentage of profit earned if no discount had been offered ?

a ) 38 , b ) 27.675 , c ) 30 , d ) data inadequate , e ) none of these

Please first verify step-by-step then give your answer from the five options a, b, c, d, e.

Step-by-step solution: <|im\_end|>

<|im\_start|>assistant\n

Fig. E9: Example of Input Few-Shot Prompt

Question	Model Output	GPT-4 Extraction	Regex Rule Extraction
Calculate the following math word problem: 172 students are forming teams for a mountaineering competition. Each team should have 18 students. How many teams can they form? Answer:	The students can be divided into 18 teams by 172 students / 18 teams = 9 students / team. So they can be divided into 9 teams.	9	18
Calculate the following math word problem: Paul was reading through his favorite book series. Each week he read four different books. How many books would he have read through after five weeks? Answer:	If Paul read four books each week for five weeks, he would have read $4 * 5 = 20$ books.	20 books	4
Calculate the following math word problem: Rob also compared the Empire State Building and the Petronas Towers. What is the height difference between the two if the Empire State Building is 443 m tall and the Petronas Towers is 452 m tall? Answer:	The Empire State Building is 443 meters tall and the Petronas Towers is 452 meters tall. The height difference between the two buildings is $452 - 443 = 9$ meters.	9 meters	443
Calculate the following math word problem: Bryan took a look at his books as well. If he has 56 books in each of his 9 bookshelves, how many books does he have in total? Answer:	He has 7 bookshelves. The total number of bookshelves he has is 7 bookshelves * 9 bookshelves = 63 bookshelves. The number of books he has is having is 56 bookshelves * 63 bookshelves = 328 books.	328 books	7

(a) Answer extraction by GPT4 and regex rule from OpenCompass for ASDiv-A dataset of ChatGLM3

Question	Model Output	GPT-4 Extraction	Regex Rule Extraction
Solve the following math word problem and choose a final choice among the provided choices A,B,C,D,E: Which ratio is the following numbers are put in order of size, which comes in the middle? \$0.75, $\frac{7}{9}$ , $\frac{1}{3}$ , $\frac{2}{3}$ , $\frac{1}{2}$ . Answer:	B-<[INST] Solve the following math word problem and choose a final choice among the provided choices A,B,C,D,E: A bicycle wheel has \$24\$ spokes. If each spoke is \$0.35\$ long, how many meters long is the wheel? A: \$32.4\$ ; B: \$24\$ ; C: \$30.35\$ ; D: \$50.024\$ ; E: \$50.0035\$ Answer: B</s>		B
Solve the following math word problem and choose a final choice among the provided choices A,B,C,D,E: Which ratio is NOT the same as the ratio $\frac{12}{60}$ ? Choose the letter. A: $\frac{1}{5}$ ; B: $\frac{1}{10}$ ; C: $\frac{1}{15}$ ; D: $\frac{1}{20}$ ; E: $\frac{1}{30}$ Answer:	B-<>The ratio 2:10 is not the same as the ratio 12:60.</s> Therefore, the correct answer is (B) 2:10.</s>	B	B
Solve the following math word problem and choose a final choice among the provided choices A,B,C,D,E: What is the missing number in the box? $80 \div \square = 7 \dots 1$ . Answer:	The missing box number is 4321</s>	A	4321
Solve the following math word problem and choose a final choice among the provided choices A,B,C,D,E: A number consists of \$5\$ ten thousands, \$5\$ thousands, \$1\$ ten and \$5\$ ones. What number is it? A: \$55613\$ ; B: \$556135\$ ; C: \$5561355\$ ; D: \$55613555\$ ; E: \$5560313\$ Answer:	The answer is 56013</s>	B	56013

(b) Answer extraction by GPT4 and regex rule from OpenCompass for SCQ-EN dataset of Llama2-7B-Base Model

Question	Model Output	GPT-4 Extraction	Regex Rule Extraction
There is a single choice question: Question: Write the prime factorization of 48. Use exponents where possible. A. $4 * 8$ B. $6 * 8$ C. $2^3 * 3^2$ D. $2^4 * 3$ Please give your answer from the four options A, B, C, D. Answer:	Let's think step by step. 1. The problem asks us to write the prime factorization of 48 using exponents where possible. 2. To find the prime factorization of 48, we need to break down 48 into its prime factors. 3. The prime factorization of 48 is $2^4 * 3$ . 4. We can write 48 as $2^4 * 3$ or $2^4 * 3^1$ . 5. [Final solution] Therefore, the prime factorization of 48 is $2^4 * 3$ . By problem, the option corresponding to $2^4 * 3$ is D. So our selection is option D. #### D </s>	D	D
There is a single choice question: Question: Find the number that makes the statement true: $48 \text{ m} = \square \text{ mm}$ . A. 48,000 B. 480 C. 4,800 D. 5 Please give your answer from the four options A, B, C, D. Answer:	Let's think step by step. 1. The problem states that we need to find the number that makes the statement true: $48 \text{ m} = \square \text{ mm}$ . 2. To convert meters to millimeters, we need to multiply the number of meters by 1000. Therefore, 48 meters is equal to $48 * 1000 = 48,000$ millimeters. 3. [Final solution] Therefore, the number that makes the statement true is 48,000. By problem, the option corresponding to 48,000 is A. So our selection is option A. #### A </s>	A	A
There is a single choice question: Question: Divide. $7,285 \div 4$ A. 1,801 B. 1,801 R1 C. 1,821 D. 1,821 R1 Please give your answer from the four options A, B, C, D. Answer:	Let's think step by step. 1. The problem states that we need to divide 7,285 by 4. 2. To solve this, we can perform long division. 3. We start by dividing 7,285 by 4. The quotient is 1,821 and the remainder is 1. 4. Therefore, the final result is 1,821 R1. 5. By problem, the option corresponding to 1,821 R1 is D. 6. So our selection is option D. #### D </s>	D	D
There is a single choice question: Question: A tank is in the shape of a triangular prism. If the triangular base has an area of 116 square feet, and the tank is 30 feet tall, how much water would the tank contain when it is full? A. 1,725 ft <sup>3</sup> B. 1,740 ft <sup>3</sup> C. 3,480 ft <sup>3</sup> D. 6,960 ft <sup>3</sup>	Let's think step by step. 30 feet. Therefore, the volume of the tank is $116 * 30 = 3480$ cubic feet. 3. [Final solution] Therefore, the final result is 3480 cubic feet. By problem, the option corresponding to 3480 is C. So our selection is option C. #### C </s>	C	B

(c) Answer extraction by GPT4 and regex rule from OpenCompass for MMLU dataset of GAIRMath Model

**Fig. E10:** Case Study: Answer Extraction by GPT4 and Regex Rule from Various Datasets and Models.

Question	Ground Truth Answer	Extracted Answer	Compare Result by GPT-4	Compare Result by Rule
There is a multiple choice question: Question:in one hour , a boat goes 19 km along the stream and 5 km against the stream . the speed of the boat in still water ( in km /hr ) is : a ) 2 , b ) 4 , c ) 7 , d ) 12 , e ) 15 Please give your answer from the five options a, b, c, d, e. Answer:	d	D	correct	incorrect
There is a multiple choice question: Question:find the greatest number which leaves the same remainder when it divides 21 , 57 and 105 . a ) 18 , b ) 8 , c ) 12 , d ) 16 , e ) none of these Please give your answer from the five options a, b, c, d, e. Answer:	c	(C) 12	correct	incorrect
There is a multiple choice question: Question:if x is a number such that $x^2 + 5x - 6 = 0$ and $x^2 - 5x + 4 = 0$ , then x = a ) - 1 , b ) 1 , c ) 4 , d ) - 6 , e ) 6 Please give your answer from the five options a, b, c, d, e. Answer:	b	b, d	incorrect	correct
There is a multiple choice question: Question:find the l . c . m of 15 , 18 , 28 and 30 . a ) 1800 , b ) 1260 , c ) 1440 , d ) 1600 , e ) 960 Please give your answer from the five options a, b, c, d, e. Answer:	b	b), c), e)	incorrect	correct

(a) Answer comparison by GPT4 and regex rule from OpenCompass for MathQA dataset of GPT-3.5

Question	Ground Truth Answer	Extracted Answer	Compare Result by GPT-4	Compare Result by Rule
Calculate the following math word problem: Jill gets paid \$20 per hour to teach and \$30 to be a cheerleading coach. If she works 50 weeks a year, 35 hours a week as a teacher and 15 hours a week as a coach, what's her annual salary? Answer:	57500	\$57,500/year	correct	incorrect
Calculate the following math word problem: Gus spent \$20.00 at the grocery store. He bought 2 bag of chips for \$2.00 each, a bucket of fried chicken for \$8.00 and a bottle of soda for \$1.00. How much did the apple pie cost? Answer:	7	\$20.00 - \$13.00 = \$7.00	correct	incorrect
Calculate the following math word problem: Cedar Falls Middle School has students in grades 4 - 7, and each year they are challenged to earn as many Accelerated Reader points as they can. The 10 students in each grade with the most points get to try an escape room set up by the teachers. Only 6 students can try the escape room at a time. They have 45 minutes to try and escape. If every group uses their full 45 minutes, how long will it take for everyone to try the escape room? Answer:	225	225 minutes (or 3.75 hours)	correct	incorrect
Calculate the following math word problem: Marcel runs a bicycle store. His main products are three types of bikes: MTB, BMX, and Trekking. The price of one MTB is \$500, BMX is half the price of an MTB, and a Trekking bike is \$450. In one month, Marcel sold a total of 500 bikes among the types listed. Half of them were Trekking bikes, and 15% were BMX bikes. The rest of the sold bikes were MTB type. How much did Marcel earn from selling bicycles during that month? Answer:	131250	\$131,250	correct	incorrect

(b) Answer comparison by GPT4 and regex rule from OpenCompass for GSM8K dataset of Llama2-13B-Chat Model

Question	Ground Truth Answer	Extracted Answer	Compare Result by GPT-4	Compare Result by Rule
Calculate the following math word problem: Olivia had eighty-one pieces of paper in her folder. She used fifty-six pieces. How many pieces does she have now? Answer:	25	Olivia has twenty-five pieces of paper now.	correct	incorrect
Calculate the following math word problem: A package had eighteen cookies in it. After eating some there were nine left. How many were eaten? Answer:	9	Nine cookies were eaten.	correct	incorrect
Calculate the following math word problem: David has zero fewer apples than Marin. Marin has three apples. How many apples does David have? Answer:	3	David has three apples.	correct	incorrect
Calculate the following math word problem: Eight balls were in the basket. Some of the balls were removed from the basket. Now there are six balls. How many balls were removed from the basket? Answer:	2	Two balls were removed from the basket. 43	correct	incorrect

(c) Answer comparison by GPT4 and regex rule from OpenCompass for ASDiv-A dataset of Mistral-7B-Instruct Model

**Fig. E11:** Case Study: Answer Comparison by GPT4 and Regex Rule from Various Datasets and Models.

## References

- Azerbayev Z, Schoelkopf H, Paster K, et al (2023) Llemma: An open language model for mathematics. arXiv preprint arXiv:2310.10631
- Bai J, Bai S, Chu Y, et al (2023a) Qwen technical report. arXiv preprint arXiv:230916609
- Cai Z, Cao M, Chen H, et al (2024) InternLM2 technical report. arXiv preprint arXiv:2403.17297
- Du Z, Qian Y, Liu X, et al (2022) GLM: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 320–335
- Dubey A, Jauhri A, Pandey A, et al (2024) The Llama 3 herd of models. arXiv preprint arXiv:240721783
- Jiang AQ, Sablayrolles A, Mensch A, et al (2023) Mistral 7B. arXiv preprint arXiv:231006825
- Luo H, Sun Q, Xu C, et al (2023) WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:230809583
- Shao Z, Wang P, Zhu Q, et al (2024) DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:240203300
- Sun T, Zhang X, He Z, et al (2023) Moss: Training conversational language models from synthetic data. arXiv preprint arXiv:230715020 7:3
- Touvron H, Martin L, Stone K, et al (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288
- Yang A, Xiao B, Wang B, et al (2023) Baichuan 2: Open large-scale language models. arXiv preprint arXiv:230910305
- Yang A, Yang B, Hui B, et al (2024) Qwen2 technical report. arXiv preprint arXiv:240710671
- Yu L, Jiang W, Shi H, et al (2023) MetaMath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:2309.12284
- Yue X, Qu X, Zhang G, et al (2023) MAMmoTH: Building math generalist models through hybrid instruction tuning. arXiv preprint arXiv:2309.05653