

MathBench: Evaluating the Theory and Application Proficiency of LLMs with a Hierarchical Mathematics Benchmark

Anonymous ACL submission

Abstract

Recent advancements in large language models (LLMs) have showcased significant improvements in mathematics. However, traditional math benchmarks like GSM8k offer a unidimensional perspective, falling short in providing a holistic assessment of the LLMs’ math capabilities. To address this gap, we introduce MathBench, a new benchmark that rigorously assesses the mathematical capabilities of large language models. MathBench spans a wide range of mathematical disciplines, offering a detailed evaluation of both theoretical understanding and practical problem-solving skills. The benchmark progresses through five distinct stages, from basic arithmetic to college mathematics, and is structured to evaluate models at various depths of knowledge. Each stage includes theoretical questions and application problems, allowing us to measure a model’s mathematical proficiency and its ability to apply concepts in practical scenarios. MathBench aims to enhance the evaluation of LLMs’ mathematical abilities, providing a nuanced view of their knowledge understanding levels and problem solving skills in a bilingual context.

1 Introduction

Mathematical reasoning and problem-solving represent pivotal facets of human intelligence and have captivated the interest of artificial intelligence (AI) research for decades. The capability of machines to grasp, interpret, and address mathematical challenges not only serves as a benchmark for their cognitive prowess but also fulfills a critical role in their deployment across various sectors.

The advent of modern Large Language Models (LLMs) such as OpenAI’s ChatGPT and GPT-4 (Achiam et al., 2023) has marked a significant milestone, showcasing an unparalleled ability to generate text that mirrors human-like discourse and to unravel intricate mathematical conundrums (Liu et al., 2023a).

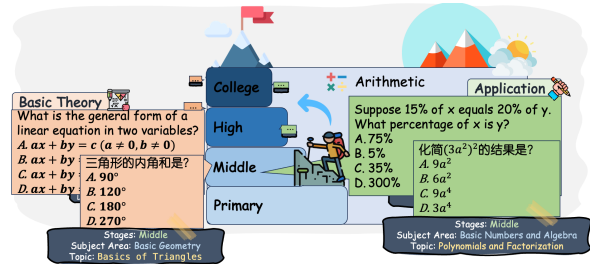


Figure 1: **MathBench Overview.** MathBench comprises multiple stages of progressively increasing challenges. Each stage encompasses bilingual theoretical and application-oriented questions, with each question precisely tagged with a three-level label to indicate its fine-grained knowledge point.

Despite these advancements, the evaluation of LLMs’ mathematical capabilities remains hampered by some inherent limitations of existing benchmarks (GSM8k (Cobbe et al., 2021), MathQA(Amini et al., 2019), etc.). These resources predominantly offer a singular perspective on problem-solving abilities and lack comprehensive difficulty grading. Math (Hendrycks et al., 2021b) attempted to classify high-school math competition problems into varying levels of complexity based on annotators’ subjective evaluations, offering an incomplete picture of mathematical proficiency. Such datasets, while valuable, fall short in encapsulating the full spectrum of mathematical knowledge and overlook the importance of fundamental theory understanding, which is essential for tackling application problems (Upadhyay and Chang, 2017a). Those limitations make it difficult to conduct a comprehensive evaluation of LLMs’ math capability (both theory and application) across different levels and disciplines and under a multilingual context.

In response to these challenges, we construct *MathBench*, a novel and comprehensive multilingual benchmark meticulously created to evaluate the mathematical capabilities of LLMs across a di-

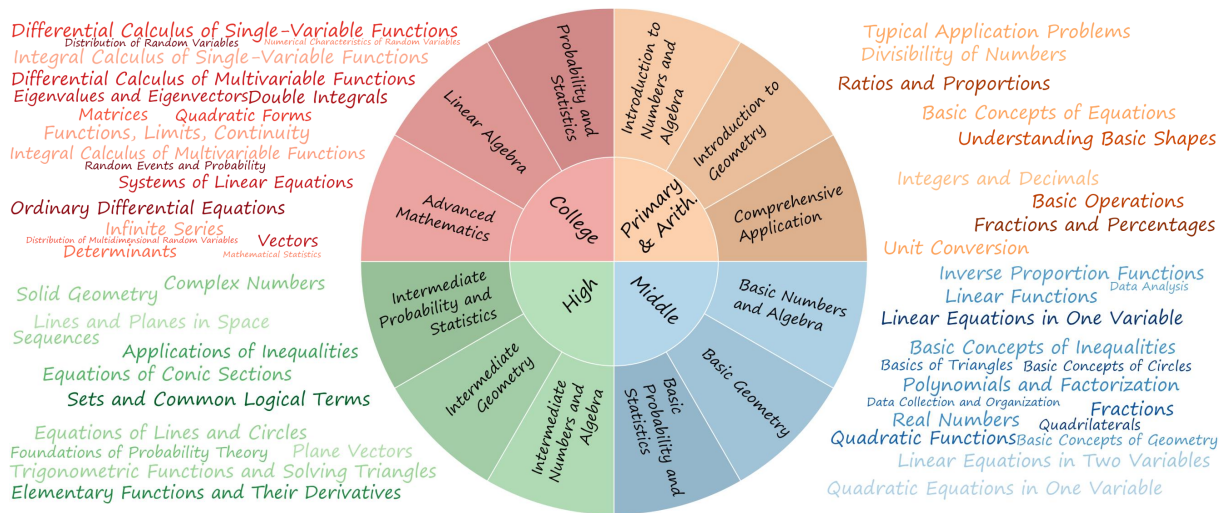


Figure 2: **Framework of MathBench**, We first categorize the mathematical content into four main educational stages and one basic arithmetic stage. Then, we extend from these to fill in two more fine-grained levels of knowledge points, forming the final MathBench framework.

verse range of difficulties, from basic arithmetic to challenging college-level mathematics. *MathBench* sets itself apart with a unique five-stage taxonomy, mapped to the educational trajectory from primary school through to college. This mechanism is designed to assess LLMs’ mathematical understanding in breadth and depth. The benchmark incorporates carefully curated questions that cover basic theory knowledge and practical applications. This dual focus enables *MathBench* to probe and interpret the models’ capabilities from a foundational standpoint. Additionally, *MathBench* supports bilingual evaluation in both Chinese and English, which facilitates a more nuanced and comprehensive assessment of LLMs’ math capabilities, offering a realistic reflection of the global landscape of mathematical knowledge.

In this paper, we detail the methodology behind the creation of MathBench, including the hierarchical knowledge system that underpins the dataset, the data collection process, and the criteria for question selection. We hope that MathBench can serve as a valuable resource for researchers and developers seeking to advance the mathematical abilities of LLMs and to understand the limitations of existing models in solving diverse and complex mathematical problems.

MathBench features the following contributions:

- We introduce *MathBench*, a comprehensive dataset that features a five-level difficulty mechanism with a hierarchical knowledge system.
- MathBench includes a wide variety of question

types, from fundamental mathematical concepts to practical application in real-world scenarios.

- We conduct extensive experiments on MathBench across different models to identify bottlenecks in current LLMs. The provided discussion and analysis are expected to offer new avenues for improving their mathematical capabilities.

2 Methodology

MathBench features a well-crafted difficulty hierarchy and an emphasis on evaluating the theoretical knowledge understanding of LLMs. Sec. 2.1 presents the tiered levels and the corresponding knowledge foundations, explaining the ability taxonomy and design rationale. Sec. 2.2 details the collection process and statistics of MathBench.

2.1 The Hierarchical Knowledge System

In MathBench, we define a knowledge framework with five main stages and three levels in order to obtain fine-grained evaluation results. Among five stages, four stages are mapped to the **four main education stages**: *Primary*, *Middle*, *High*, and *College*, while the other stage is named *Arithmetic*, serving as the foundation of the remaining four stages.¹ Each **Stage** in MathBench is associated with two fine-grained knowledge levels: **Subject Area** and **Topic**, accordingly. As shown in Figure 2, we extend MathBench from the basic stages to a comprehensive range of mathematical concepts and

¹The ‘Arithmetic’ stage evaluates the ability to perform four basic math operations: add, subtract, multiply, divide.

Table 1: **Overview of Datasets Included in MathBench.** MCQ stands for Multi-Choice Question.

Name	Dataset Type	Question Type
GSM-X-CN	Self-Collected	Open-ended QA
GSM-X-Plus	Self-Collected	Open-ended QA
CEVAL-Math	Open Source	MCQ
MMLU-College-Math	Open Source	MCQ
Math401	Open Source	MCQ
Hungarian-Math-MCQ	Self-Collected	MCQ
AMC-8 & 12	Self-Collected	MCQ
SAT	Self-Collected	MCQ
Gaokao	Self-Collected	MCQ
Zhongkao	Self-Collected	MCQ
Kaoyan	Self-Collected	MCQ
SciBench	Open Source	MCQ
Arithmetic-HG	Open Source	Open-ended QA
Theory-Knowledge-Primary	Self-Collected	MCQ
Theory-Knowledge-Middle	Self-Collected	MCQ
Theory-Knowledge-High	Self-Collected	MCQ
Theory-Knowledge-College	Self-Collected	MCQ

problem-solving skills. Such taxonomy is designed to capture the depth and breadth of mathematical knowledge, from foundational arithmetic to complex, abstract college-level concepts.

Subject Areas include major mathematical disciplines such as Algebra, Geometry, Trigonometry, Calculus, Statistics, Probability, *etc.*. This categorization allows for a wide range of questions, facilitating an organized approach to covering the diverse areas of mathematics. Within each subject area, we further refine the classification into specific **Topics**. For example, under Algebra, topics might include Linear Equations, Quadratic Equations, Polynomials, and Functions. The Topic-level granularity ensures that the dataset can provide detailed insights into a model’s understanding and proficiency in specific areas of mathematics.

In MathBench, each question is tagged with metadata indicating its stage (Primary, Middle, High, College, or Arithmetic), subject area, and topic. Such tags enable a fine-grained analysis of models’ performance across different areas of mathematics and allow researchers to identify specific strengths and weaknesses in mathematical understanding.

Moreover, the inclusion of the Arithmetic stage emphasizes the importance of mastering basic math operations, which is the foundation of all subsequent mathematical learning and problem-solving.

2.2 Data Collection and Statistics

With the pre-defined knowledge framework, we primarily collect questions from two perspectives: (a). *theoretical knowledge questions*, to test the model’s

grasp of basic formulas, theories, and their corollaries, which are the foundation for solving mathematical problems; (b). *practical application questions*, which often require a good understanding of the fundamental theories, reflecting the ability to apply these theories in practice.

Question Format Definition. During the evaluation, some models struggle with open-ended questions and fail to follow instructions and provide plain and concise answers. Therefore, we reformulate questions that could have complex answers² into the multiple-choice format, typically with four options. During collection and annotation, we ensure the uniqueness of the correct answer and the high confusion-level of distractive options.

Theoretical Knowledge Questions. For theoretical knowledge questions, we collect the definition and detailed corollaries of knowledge points topic by topic from the math textbooks and the Internet. We then transform them to multi-choice questions with high-quality annotations.

Practical Application Questions. On selecting the practical application questions, we primarily consider the following aspects: 1. The question needs to match the corresponding education level; 2. The questions should comprehensively cover the previously defined knowledge taxonomy; 3. The questions should be well-formulated so that LLMs can answer them properly. We primarily focus on stage-based educational exams or competitions. Those questions are comprehensive and representative, offering a certain degree of difficulty gradient, such as ZhongKao, GaoKao in Chinese Math and AMC, SAT in English math. Additionally, we incorporate open-source questions to enhance the diversity and breadth of the questions. We list the sources of questions in MathBench in Table 1.

Quality Screening. To enhance the quality of the MathBench dataset, we implement a semi-automated question filtering process to mitigate issues such as intrinsic question errors and alignment with educational stages utilizing GPT-4, details presented in Appendix A.3.

Dataset Summary. We curate 3709 questions for the final MathBench, including both Chinese and English languages across five stages with three-level knowledge taxonomy. MathBench includes

²All theoretical knowledge questions and practical application questions from middle school to college level

208 2209 theoretical questions and 1,500 practical ap- 253
209 plication questions, all of which have undergone 254
210 semi-automated screening. Detailed statistics can 255
211 be found in the Appendix A.1. 256

212 3 Experiments and Analysis 257

213 3.1 Configuration 258

214 **Evaluation Protocols.** We employ CircularEval 259
215 (CE) (Liu et al., 2023b) as our principal evaluation 260
216 methodology. CE systematically assesses an N - 261
217 option multi-choice question by evaluating it N 262
218 times, each time permuting the order of the options. 263
219 To ensure uniformity across evaluations, we set the 264
220 maximum output length at 512 tokens and use the 265
221 greedy decoding strategy for all LLMs. We adopt 266
222 the few-shot setting for open-ended questions and 267
223 the zero-shot setting for multi-choice questions. 268

224 **Evaluated Models.** Our evaluation encom- 271
225 passes both API-based commercial LLMs and 272
226 open-source LLMs, covering a total of 20 mod- 273
227 els. Based on MathBench, we deliver a thorough 274
228 evaluation of the capabilities of current LLMs. We 275
229 list all evaluated LLMs below: 276

- 230 • API models: OpenAI GPT-3.5 and GPT-4³. 277
- 231 • OpenSource LLMs: We evaluate a wide spec- 278
232 trum of LLMs, including QWen (Bai et al., 2023), 279
233 InternLM (Team, 2023), Yi⁴, Baichuan2 (Yang 280
234 et al., 2023), DeepSeek(DeepSeek-AI et al., 2024) 281
235 and ChatGLM3 (Zeng et al., 2022). 282
- 236 • OpenSource Math LLMs: MetaMath-llemma(Yu 283
237 et al., 2023), DeepSeekMath(Shao et al., 2024), 284
238 MAMmoTH(Yue et al., 2023) and InternLM- 285
239 Math(Ying et al., 2024). 286

240 3.2 Main Results 287

241 The overall experimental results are shown in Ta- 288
242 ble 2. We report the average score of theoret- 289
243 ical and application questions for all stages expect 290
244 *Arithmetic*, which only has application questions. 291

245 Among all evaluated models, GPT-4 consistently 292
246 outperforms the others, showcasing superior per- 293
247 formance across all metrics. The second best LLM 294
248 is Qwen-72B, its outstanding performance distin- 295
249 guishes itself as the leading player among all open- 296
250 source models. We also notice that DeepSeek- 297
251 Math-7B-RL, an LLM dedicated to mathematical 298
252 tasks, secures its position as the second-best open- 299

source model in mathematics, which is impressive 300
given its small parameter size. 301

Among Open-Source Chat Models, perfor- 302
mances across models with $\sim 7B$, $\sim 20B$, and $\sim 70B$ 303
parameter size reveal distinct capabilities: 304

$\sim 7B$ Chat Models. InternLM2-Chat-7B emerges 305
as the superior model at the $\sim 7B$ scale and outper- 306
forms other 7B Chat models across all stages. It’s 307
noteworthy that, as the difficulty of problems in- 308
creases, the gap between InternLM2-Chat-7B and 309
other models also grows. For instance, on the five 310
stages from *Arithmetic* to *College Math*, it outper- 311
forms ChatGLM3-6B by 29%, 67%, 92%, 157%, 312
and 258%, respectively. The trend indicates that 313
as the difficulty escalates, the performance dispar- 314
ity between models significantly increases since 315
higher-stage math problems often involve more 316
complex concepts and problem-solving strategies, 317
imposing greater demands on the models’ compre- 318
hension and reasoning abilities. All $\sim 7B$ models 319
struggle with advanced mathematical problems, in- 320
dicating a challenge in smoothly resolving complex 321
questions for small-scale LLMs. 322

$\sim 20B$ Chat Models. Qwen-14B-Chat performs 323
the best at the $\sim 20B$ scale, followed by InternLM2- 324
Chat-20B. Though Yi-34B-Chat has a much larger 325
parameter size, it lags behind other $\sim 20B$ models. 326
Similar to $\sim 7B$ models, models around $\sim 20B$ also 327
struggle with more complex mathematical prob- 328
lems at the *High School* and *College* stage. 329

$\sim 70B$ Chat Models and Math Models. Large- 330
scale Open-Source LLMs demonstrate far bet- 331
ter performance compared to their small/medium- 332
scale counterparts. Qwen-72B-Chat, for instance, 333
achieves excellent results across all stages, which is 334
comparable to the state-of-the-art GPT-4. Among 335
Math LLMs, DeepSeek-Math-7B-RL excels in 336
both basic *Arithmetic* and *College* math, outper- 337
forming not only its peers but also the much heavier 338
DeepSeek-67B-Chat. 339

330 3.3 Detailed Analysis 340

341 With MathBench, we can easily assess the model’s 341
mathematical capabilities at different granularities 342
including education stage, language, subject area, 343
or even specific topics with questions on both theo- 344
ries and applications. Below, we will delve deeper 345
into the evaluation results and discuss about the 346
following questions: 347

**How Models’ Scores on Application Problems 348
Vary Across Stages?** Figure 3 presents the aver- 349

³GPT-4 version: gpt-4-0125-preview; GPT-3.5 ver- 350
sion: gpt-3.5-turbo-0125 351

⁴<https://github.com/01-ai/Yi> 352

Table 2: **Overall Comparison of Models on MathBench.** Models are classified into three categories according to their purpose and origin. The model name in **bold** indicates the top performer among Open-Source or API models, while an underline signifies the leading model within a similar parameter size group.

Models	Arithmetic	Primary	Middle	High	College	Average
★API Models						
GPT-3.5	70.3	67.9	39.3	30.6	32.2	48.1
GPT-4	76.3	82.9	69.8	56.6	59.0	68.9
♡Open-Source Chat Models						
ChatGLM3-6B	41.0	40.5	21.4	11.5	6.3	24.1
Yi-6B-Chat	35.7	41.1	20.3	11.5	9.1	23.5
InternLM2-Chat-7B	<u>53.0</u>	<u>67.5</u>	<u>41.0</u>	<u>29.6</u>	<u>22.6</u>	<u>42.7</u>
Qwen-7B-Chat	51.3	50.2	32.6	20.2	17.3	34.3
Deepseek-7B-Chat	46.0	39.3	15.5	9.6	9.2	23.9
Baichuan2-13B-Chat	46.0	54.2	29.5	16.6	14.3	32.1
Qwen-14B-Chat	<u>64.7</u>	66.1	<u>49.2</u>	32.8	<u>27.2</u>	<u>48.0</u>
InternLM2-Chat-20B	62.7	<u>70.0</u>	47.4	<u>33.7</u>	23.3	47.4
Yi-34B-Chat	51.0	64.8	38.0	23.2	17.8	39.0
Deepseek-67B-Chat	61.3	77.2	48.4	36.3	36.8	52.1
Qwen-72B-Chat	72.0	80.1	64.8	47.8	40.8	61.1
△Mathematical Models						
MammoTH-7B	26.7	18.1	5.3	4.8	3.7	11.7
Metamath-Llemma-7B	48.7	35.3	16.1	15.5	10.1	25.1
InternLM2-Chat-Math-7B	53.7	66.0	49.0	34.3	26.9	46.0
Deepseek-Math-7B-Instruct	61.0	73.7	42.2	34.9	29.9	48.3
Deepseek-Math-7B-RL	<u>67.7</u>	80.8	<u>57.2</u>	<u>45.4</u>	42.7	<u>58.8</u>
MammoTH-13B	35.0	34.8	10.7	9.9	10.6	20.2
InternLM2-Chat-Math-20B	58.7	71.1	55.5	41.8	31.9	51.8
MammoTH-70B	35.7	59.3	28.1	23.6	24.5	34.2

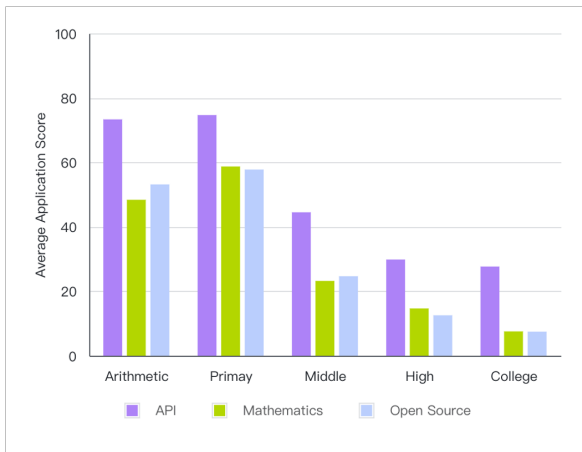


Figure 3: **Scores of Application Problems at Each Stage.** Models exhibit similar performances in *Arithmetic* and *Primary* stages, while demonstrating a clear performance decline from *Primary* to *College* stages.

age performance of all aforementioned models on application questions in MathBench. Most models perform reasonably well on *Arithmetic* and *Primary* math problems. However, their effectiveness drastically declines when it comes to the *Middle*

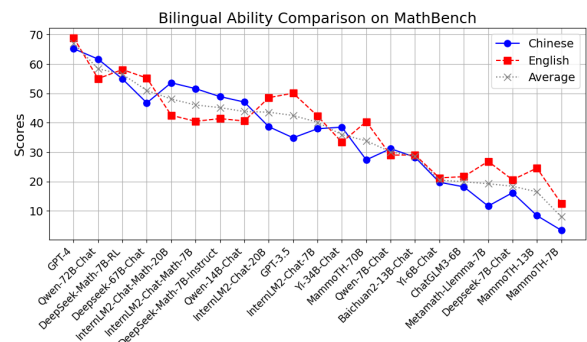


Figure 4: **Bilingual Comparison on MathBench.** showcasing scores in Chinese, English, and their average for the gray dashed line. The *Arithmetic* stage is not include because there no impact of language in it.

stage or above. Such phenomenon suggests that existing models are good at tasks that can be solved through direct computation, pattern recognition, or memorizing basic concepts. However, they showcase inferior performance when solving more complex math problems.

Is There A Gap between Theory Understanding and Application Capabilities? Theories serve

308
309
310
311
312
313
314
315

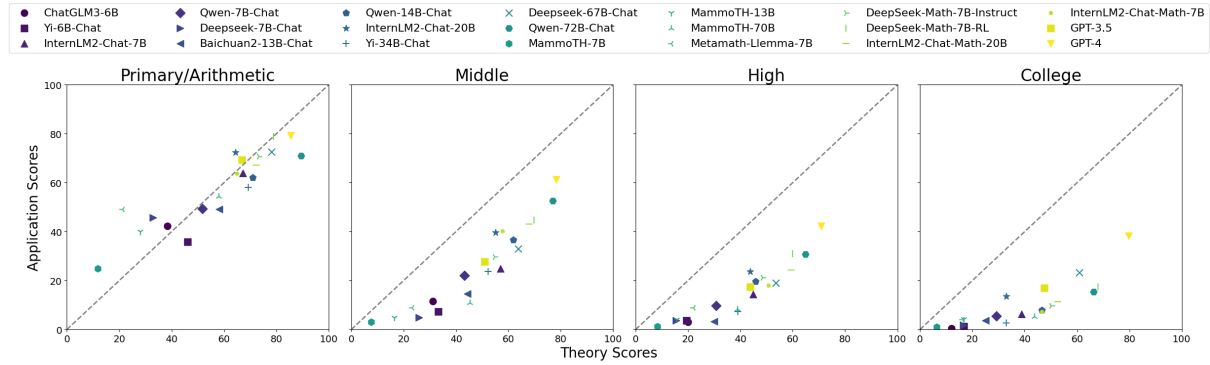


Figure 5: **Theoretical Score vs. Applied Score on MathBench.** *Primary* and *Arithmetic* are averaged because they share the same theory knowledge points.

as the foundation for addressing the majority of application problems. As illustrated in Figure 5, we present the trend of LLMs in terms of theoretical and application scores across different stages. In the *Primary* stage, the two scores are highly correlated for most LLMs, with only a few exceptions. Among top-ranked models, Qwen-72B-Chat demonstrates the best theoretical ability, while GPT-4 demonstrates superior application ability. When it comes to more advanced stages, models require better computational and reasoning capabilities to achieve good application scores. GPT-4 leads in the application track across all stages, while the gap is larger in more advanced stages. For example, comparing to Qwen-72B-Chat, the difference in theoretical and application scores (D_t, D_a) increases from (1.4, 8.7) in the *Middle* stage to (6.0, 11.7) in the *High* stage, and finally to (13.5, 23.0) in the *College* stage. Moreover, from the *Middle* stage onwards, there is a general trend of decline in both theoretical and application abilities of models. Compared to theoretical scores, the decline in application scores is more serious.

Which Model Performs Better under the Bilingual Scenario? Figure 4 demonstrates the bilingual capabilities of various LLMs on MathBench, indicating the importance of linguistic versatility in mathematical tasks that demand an understanding of nuances in language and math concepts across different languages. Among all LLMs, GPT-4 leads with the highest bilingual score of 67.1, showing a balanced performance between Chinese (65.2) and English (69.0). This demonstrates GPT-4’s advanced bilingual processing abilities. Other models including Qwen-72B-Chat and DeepSeek-Math-7B-RL also exhibit significant bilingual capabilities. It’s also noteworthy that among all LLMs

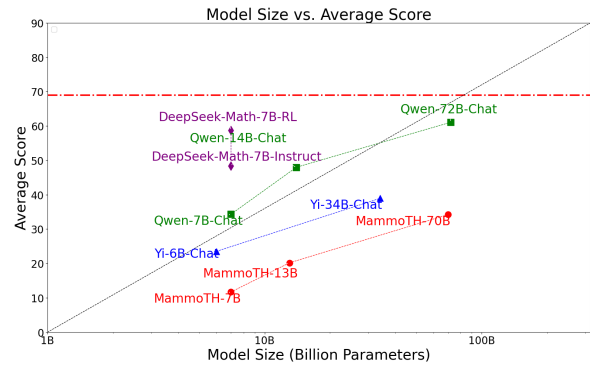


Figure 6: **Model Size vs. Average Score.** The comparison chart of model parameter size versus performance on MathBench for selected representative models, with models from the same series connected by lines of the same color. The horizontal red dotted line represents the score of GPT-4.

evaluated, most of them feature a much larger performance gap between Chinese and English, compared to GPT-4.

4 Discussion

4.1 Effect of Model Size on Math Capabilities

We found that for models of different sizes within the same series, most of them conform to the Scaling Law (Kaplan et al., 2020) on MathBench. For example, Qwen series, MammoTH series, and Yi series have shown steady improvement in their MathBench scores as the parameter size increases, as shown in Figure 6. However, it doesn’t mean that models with small parameter sizes can not achieve good math performance. For instance, DeepSeek-Math-7B demonstrates outstanding performance on MathBench and outperforms models with 10x parameters, including DeepSeek-72B and a larger math model MammoTH-70B.

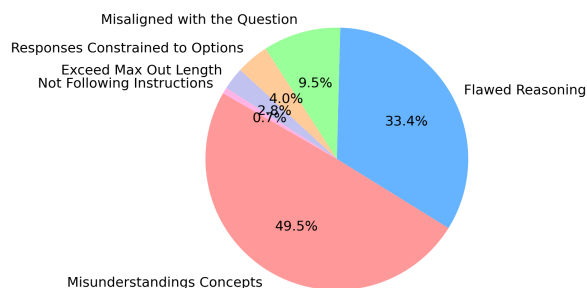


Figure 7: **Response Error Analysis for Both Theoretical and Application Questions.** The predominant sources of errors are a fundamental misunderstanding of the concepts, followed by incorrect reasoning paths.

4.2 Error analysis

In our study, we conduct a comprehensive error analysis on a set of 80 theoretical and 100 application questions random selected from every stages, for models selected across different scales, as illustrated in Figure 7. The error categories are uniformly observed across all evaluated models, indicating common challenges that transcend specific parameter scales. Our selection of models includes GPT-3.5, GPT-4, InternLM-Chat-7B, Qwen-14B-Chat, Qwen-72B-Chat, Deepseek-Math-7B-RL and MammoTH-70B. Detailed cases for error analysis can be found in Appendix C.2.

Insufficiency of knowledge. For theoretical questions, 78% of model errors are due to misconceptions about mathematical concepts, which notably emerged as a significant concern in several models. Such errors accounted for 49.5% of all mistakes, underscoring a general challenge in grasping fundamental knowledge and terminology.

Deficiencies in reasoning. Furthermore, models exhibited shortcomings in logical reasoning, with 33.4% of errors attributed to logically consistent but flawed reasoning processes. Moreover, errors such as reasoning that deviated from the intended query accounting for 9.6%, underscored the models' limitations in understanding user intentions and providing pertinent responses. We also notice that errors related to reasoning increased with task difficulty.

Response length limit. Though statistically not the primary error mode (4.0%), responses that exceeded the token limit shed light on the challenge of reasoning complex tasks within limited length and adhering to given instructions.

Other cases. Occasionally, models will generate responses devoid of an explicit reasoning process,

obstructing additional scrutiny. Moreover, models endowed with enhanced reasoning capabilities exhibit a greater capacity for critical thinking regarding the options presented, thereby offering alternative answers that transcend the limitations of predetermined choices.

4.3 Reasoning Path

Analyzing the reasoning paths of various models across multiple difficulty levels reveals significant performance disparities. We set a brief discussion below and provide more detailed cases for reasoning path analysis in Appendix C.3.

Performance across diverse difficulties. In straightforward scenarios, models swiftly solve the problems with direct reasoning and yield logical outcomes. Yet, complex issues, marked by dense symbols, vast knowledge, and intricate links, necessitate broader knowledge navigation, accentuating divergences in deductive strategies.

Reasoning paths of chat models with different parameter sizes. Small-scale chat models strive for logical coherence in mathematics, yet may make mistakes due to knowledge deficiencies, particularly in symbol interpretation and relational understanding. In contrast, large-scale models feature expansive knowledge and nuanced insights, which enhance symbol processing and minimizing knowledge gaps. However, even with substantial parameters, challenges in efficient knowledge management persist, occasionally leading to irrelevant diversions and diminished reasoning efficacy.

Reasoning paths of math models. Specialized math models, despite the smaller parameter sizes, exhibit superior mathematical comprehension and systematic logical reasoning. They excel in applying mathematical knowledge and notation to reason through complex problems.

Superlative deductive navigation of API models. GPT-4 stands out for its effective reasoning and deep problem comprehension. It engages in logical, coherent, and succinct discussions, adeptly navigate complex reasoning paths, and manage mathematical symbols effectively. GPT-4 distinctively recognizes problem statement ambiguities, showcasing a detailed and nuanced reasoning process.

5 Related works

Solving math word problems through automated methods has been a long-standing concern for researchers. This section summarizes seminal studies

457	and delineates key evaluation datasets proposed	Integrating mathematical problems with do-	506
458	for assessing mathematical problem-solving ap-	main knowledge NumGLUE (Mishra et al.,	507
459	proaches, tracing the field’s evolution from its ori-	2022) not only assesses the ability of models to	508
460	gins to the present day.	solve mathematical problems given direct compu-	509
		tational expressions, but it also designs multiple	510
461	Preliminary Mathematical Datasets Previous	tasks to comprehensively evaluate the models’ abil-	511
462	works proposed datasets such as Alg514 (Kush-	ities to use other reasoning skills, such as common	512
463	man et al., 2014), SingleEq (Koncel-Kedziorski	sense and reading comprehension. Lila (Mishra	513
464	et al., 2015), and DRAW-1K (Upadhyay and Chang,	et al., 2023) is developed through the extension of	514
465	2017b) are primarily concentrated on elementary	20 datasets that cover a broad range of mathemat-	515
466	linear algebraic problems. Similarly, datasets like	ical topics. This dataset exhibits varying degrees	516
467	AddSub (Hosseini et al., 2014) and SingleOp (Roy	of linguistic complexity and features diverse ques-	517
468	et al., 2015) are exclusively dedicated to funda-	tion formats as well as background knowledge re-	518
469	mental arithmetic operations: addition, subtraction,	quirements. These works inspire us to design more	519
470	multiplication, and division. These datasets are	diversified testing scenarios.	520
471	very limited both in the form and content of their		
472	assessments, focusing solely on a specific small	6 Conclusion	521
473	part of basic mathematics.		
		In summary, MathBench adopts structured ap-	522
474	Benchmarks tailored to specific educational	proaches to categorize questions by stage and	523
475	tiers Some benchmarks are designed based on	knowledge level. It aims to provide a comprehen-	524
476	educational levels. Math23k (Wang et al., 2017)	sive evaluation of LLMs’ mathematical proficiency.	525
477	collects a corpus of real math word problems for	By covering a wide range of subject areas and top-	526
478	elementary school students. While ASDiv (Miao	ics across educational stages, MathBench offers	527
479	et al., 2021) expands the textual patterns to encom-	a unique resource for researchers and educators	528
480	pass most problem types found in elementary math-	interested in advancing the field of mathematical	529
481	ematics. GSM8K (Cobbe et al., 2021) presents a	learning and assessment.	530
482	high-quality collection of elementary mathematical		
483	word problems that, on average, require multiple	7 Limitations	531
484	steps to solve and provide solutions in natural lan-		
485	guage annotations. These datasets mostly focus	We have developed a comprehensive mathemati-	532
486	on elementary mathematics and seldom examine	cal evaluation benchmark, MathBench, which in-	533
487	college-level knowledge.	cludes a detailed knowledge framework and multi-	534
		dimensional, fine-grained mathematical questions.	535
488	Enriching the diversity of mathematical prob-	Despite its strengths, the benchmark currently has	536
489	lem types within benchmarks MathQA (Amini	several limitations, which are summarized as fol-	537
490	et al., 2019) seeks to categorize problems from	lows:	538
491	AQuA (Ling et al., 2017) into different mathemat-	Data Source: To enhance diversity, some ques-	539
492	ical domains based on the frequency of mathemat-	tions were sourced from open-source datasets(~	540
493	ical terminology used. Mathematics Dataset (Sax-	19%). However, these open-source questions may	541
494	ton et al., 2019) expands the subject of mathematics	be subject to data contamination, which could com-	542
495	and this dataset covers a broader spectrum of math-	promise the assurance that models have not been	543
496	ematics, including arithmetic, algebra, probability,	exposed to these questions before. In future iter-	544
497	and calculus. MATH (Hendrycks et al., 2021b) fea-	ations, we plan to automate the construction of	545
498	tures a higher level of complexity, comprising prob-	questions across various stages to more effectively	546
499	lems ranging from arithmetic to calculus, and aims	test the models’ genuine mathematical capabilities.	547
500	at testing models’ capabilities in understanding and	Lack of Detailed Reasoning Paths: Given the	548
501	solving complex mathematical challenges. While	diversity of questions and time constraints, Math-	549
502	these efforts have enhanced the diversity of the data	Bench currently does not provide detailed reason-	550
503	in certain aspects, they are lacking in diversity in	ing paths for each question. This limitation makes	551
504	other aspects such as question formulation (Saxton	it challenging to unlock the full potential of the	552
505	et al., 2019).	questions. Moving forward, we aim to investigate	553
		semi-automated methods to offer both natural lan-	554

555	guage and code-based reasoning approaches for		
556	each question, thereby maximizing the value of		
557	MathBench’s questions.		
558	8 Ethical Considerations		
559	For our benchmarks, we relied on reference ma-		
560	terials and APIs that are accessible to the public,		
561	thereby avoiding any potential harm to individuals		
562	or groups. The data produced by the LLMs under-		
563	went a meticulous human selection and processing		
564	phase to ensure the protection of privacy and con-		
565	fidentiality. We did not use any personally identi-		
566	fiable information, and all data were anonymized		
567	prior to analysis. Additionally, we employed Chat-		
568	GPT and Grammarly to refine our manuscript’s		
569	language.		
570	References		
571	Josh Achiam, Steven Adler, Sandhini Agarwal, Lama		
572	Ahmad, Ilge Akkaya, Florencia Leoni Aleman,		
573	Diogo Almeida, Janko Altenschmidt, Sam Altman,		
574	Shyamal Anadkat, et al. 2023. Gpt-4 technical report.		
575	<i>arXiv preprint arXiv:2303.08774</i> .		
576	Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-		
577	Kedziorski, Yejin Choi, and Hannaneh Hajishirzi.		
578	2019. Mathqa: Towards interpretable math word		
579	problem solving with operation-based formalisms .		
580	Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang,		
581	Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei		
582	Huang, et al. 2023. Qwen technical report. <i>arXiv</i>		
583	<i>preprint arXiv:2309.16609</i> .		
584	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,		
585	Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias		
586	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro		
587	Nakano, Christopher Hesse, and John Schulman.		
588	2021. Training verifiers to solve math word prob-		
589	lems . <i>CoRR</i> , abs/2110.14168.		
590	DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting		
591	Chen, Shanhuang Chen, Damai Dai, Chengqi Deng,		
592	Honghui Ding, Kai Dong, Qiu Shi Du, Zhe Fu,		
593	Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge,		
594	Kang Guan, Daya Guo, Jianzhong Guo, Guangbo		
595	Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan		
596	Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li,		
597	Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu,		
598	Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan		
599	Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma,		
600	Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu,		
601	Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli		
602	Sha, Zhihong Shao, Junxiao Song, Xuecheng Su,		
603	Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingx-		
604	uan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang,		
605	Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie,		
606	Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu,		
607	Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping		
	Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong	608	
	Zhang, Liyue Zhang, Mingchuan Zhang, Minghua	609	
	Zhang, Wentao Zhang, Yichao Zhang, Chenggang	610	
	Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou,	611	
	Qihao Zhu, and Yuheng Zou. 2024. Deepseek llm:	612	
	Scaling open-source language models with longter-	613	
	mism .	614	
	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	615	
	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	616	
	2021a. Measuring massive multitask language under-	617	
	standing .	618	
	Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul	619	
	Arora, Steven Basart, Eric Tang, Dawn Song, and	620	
	Jacob Steinhardt. 2021b. Measuring mathematical	621	
	problem solving with the math dataset . <i>Cornell Uni-</i>	622	
	<i>versity - arXiv</i> , <i>Cornell University - arXiv</i> .	623	
	Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren	624	
	Etzioni, and Nate Kushman. 2014. Learning to solve	625	
	arithmetic word problems with verb categorization .	626	
	In <i>Proceedings of the 2014 Conference on Empirical</i>	627	
	<i>Methods in Natural Language Processing (EMNLP)</i> .	628	
	Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei	629	
	Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu,	630	
	Chuan Cheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu,	631	
	Maosong Sun, and Junxian He. 2023. C-eval: A	632	
	multi-level multi-discipline chinese evaluation suite	633	
	for foundation models .	634	
	Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B.	635	
	Brown, Benjamin Chess, Rewon Child, Scott Gray,	636	
	Alec Radford, Jeffrey Wu, and Dario Amodei. 2020.	637	
	Scaling laws for neural language models .	638	
	Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish	639	
	Sabharwal, Oren Etzioni, and Siena Dumas Ang.	640	
	2015. Parsing algebraic word problems into equa-	641	
	tions . <i>Transactions of the Association for Computa-</i>	642	
	<i>tional Linguistics</i> , 3:585–597.	643	
	Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and	644	
	Regina Barzilay. 2014. Learning to automatically	645	
	solve algebra word problems . In <i>Proceedings of the</i>	646	
	<i>52nd Annual Meeting of the Association for Computa-</i>	647	
	<i>tional Linguistics (Volume 1: Long Papers)</i> , pages	648	
	271–281.	649	
	Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blun-	650	
	som. 2017. Program induction by rationale genera-	651	
	tion: Learning to solve and explain algebraic word	652	
	problems . <i>arXiv preprint arXiv:1705.04146</i> .	653	
	Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding,	654	
	Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen,	655	
	Bo Jiang, Aimin Zhou, and Liang He. 2023a. Mathe-	656	
	matical language models: A survey .	657	
	Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li,	658	
	Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi	659	
	Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua	660	
	Lin. 2023b. Mmbench: Is your multi-modal model	661	
	an all-around player?	662	

663	Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers.	717
664		718
665		719
666	Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Taffjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2023. Lila: A unified benchmark for mathematical reasoning.	720
667		721
668		722
669		723
670		724
671	Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks.	725
672		726
673		727
674		
675	Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about Quantities in Natural Language. <i>Transactions of the Association for Computational Linguistics</i> , 3:1–13.	728
676		729
677		730
678		731
679	David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models.	732
680		733
681		734
682	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.	735
683		736
684		
685		
686		
687	InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.	737
688		
689		
690	Shyam Upadhyay and Ming-Wei Chang. 2017a. Annotating derivations: A new evaluation strategy and dataset for algebra word problems.	738
691		
692		
693	Shyam Upadhyay and Ming-Wei Chang. 2017b. Annotating derivations: A new evaluation strategy and dataset for algebra word problems.	739
694		
695		
696	Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.	740
697		
698		
699		
700		
701	Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In <i>Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing</i> , pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.	741
702		
703		
704		
705		
706		
707	Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. <i>arXiv preprint arXiv:2309.10305</i> .	742
708		
709		
710		
711		
712	Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu	743
713		
714		
715		
716		
	Wang, Kai Chen, and Dahua Lin. 2024. Internlm-math: Open math large language models toward verifiable reasoning.	744
		745
		746
		747
		748
		749
		750
		751
		752
	Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Meta-math: Bootstrap your own mathematical questions for large language models.	
	Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks?	
	Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning.	
	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. <i>arXiv preprint arXiv:2210.02414</i> .	

A MathBench Statistics

A.1 Dataset Statistics

The detailed statistics of MathBench questions, Table 3 for the data distribution of theoretical and application questions across various stages, Table 4 for fine-grained knowledge levels.

Table 3: Detailed Composition of the MathBench

Stage	Type	English	Chinese	Total
Arithmetic	Theoretical	-	-	-
	Application	300	-	300
Primary	Theoretical	109	208	317
	Application	150	150	300
Middle	Theoretical	175	334	509
	Application	150	150	300
High	Theoretical	281	470	751
	Application	150	150	300
College	Theoretical	316	316	632
	Application	150	150	300

A.2 Data collection details

For self-collected questions in MathBench, We primarily collect through the following methods:

For the Primary stage GSM-X-CN and GSM-X-Plus datasets, we semi-automatically generate new questions using GPT-4. Specifically, the construction of the GSM-X-CN Chinese question set involved two steps:

We first translate English questions in GSM8k test set into Chinese using GPT-4, resulting in a

Table 4: **MathBench Subject Area Statistics**. Data is shown at the Subject Area level for conciseness, omitting the more detailed Topic level due to its breadth.

Stage	Subject Area	English	Chinese	Total
Primary & Arith.	Introduction to Numbers and Algebra	44	73	117
	Introduction to Geometry	10	62	72
	Comprehensive Application	55	73	128
Middle	Basic Numbers and Algebra	133	182	315
	Basic Geometry	33	137	170
	Basic Probability and Statistics	9	15	24
High	Intermediate Numbers and Algebra	146	189	335
	Intermediate Geometry	114	219	333
	Intermediate Probability and Statistics	21	62	83
College	Advanced Mathematics	119	119	238
	Linear Algebra	99	99	198
	Probability and Statistics	98	98	196

Chinese version of GSM8k. We then replace the entity names under the Chinese context while ensuring that the questions’ meanings remained unchanged. This process creates elementary-level questions suitable for Chinese Q&A.

For the GSM-X-Plus dataset, which is in English, we generate new questions by first generating solution code for the original test set questions. We then replace some numeric parameters (taken from the original questions) in the question with multiples of the variable k . By executing the modified solution code, we obtain the new answers. In MathBench, we set $k \in (2, 10)$.

For exams such as AMC, GaoKao, ZhongKao, *etc.*, we initially collect relevant questions from the Internet. These questions are then underwent processing and annotation by domain experts. Questions for primary and secondary education levels are handled and annotated by undergraduate students, while questions for university-level exams were processed and annotated by graduate students specializing in mathematics or computer science. The description of knowledge-based questions is provided in Sec. 2.2.

In addition to the self-collected datasets described above, we also incorporate questions from the following open-source datasets: CEVAL (Huang et al., 2023), MMLU (Hendrycks et al., 2021a), Arithmetic-HG, Math401 (Yuan et al., 2023) and SciBench (Wang et al., 2024). All open-source datasets we used are MIT License.

A.3 Quality Screening

Given the wide variety of sources and types of questions, we notice that the following issues may affect the benchmark quality: 1. Intrinsic errors in the questions, such as being unanswerable or having multiple correct answers. 2. Questions of low evaluation value, too difficult or too trivial for the intended education stage.

All of the above situations can easily lead to unstable model responses and increased probability of incorrect answers in CircularEval. To address these issues, we employ a novel semi-automated question filtering approach for quality screening.

Specifically, we use GPT-4 to perform Circular Evaluation (CE) on all questions. We then select questions that GPT-4 answered incorrectly 0, 1, or 2 times out of four attempts ($CE = 0$, $CE = 1$, $CE = 2$) for manual review to ensure the overall question quality.

B Detailed Experimental Results

B.1 THEORY AND APPLICATION

The corresponding results is presented in Table 5.

B.2 BILINGUAL

The corresponding prompt is presented in Table 6.

Table 5: Detailed Results of Theory and Application Score on MathBench (Theory/Application)

Models	Primary	Middle	High	College
★API Models				
GPT-3.5	66.8/69.0	50.9/27.7	43.8/17.3	47.5/17.0
GPT4	85.4/80.3	78.3/61.3	70.9/42.3	79.8/38.3
♡Open-Source Chat Models				
ChatGLM3-6B	38.4/42.7	31.2/11.7	20.0/3.0	12.0/0.5
Yi-6B-Chat	46.2/36.0	33.3/7.3	19.4/3.7	16.8/1.3
InternLM2-Chat-7B	67.3/67.7	57.0/25.0	44.9/14.3	38.8/6.3
Qwen-7B-Chat	51.7/48.7	43.3/22.0	30.7/9.7	29.1/5.5
Deepseek-7B-Chat	33.0/45.7	26.0/5.0	15.4/3.7	16.6/1.8
Baichuan2-13B-Chat	58.1/50.3	44.3/14.7	29.9/3.3	25.0/3.7
Qwen-14B-Chat	70.9/61.3	61.8/36.7	45.9/19.7	46.5/7.8
InternLM2-Chat-20B	64.3/75.7	55.2/39.7	43.8/23.7	32.9/13.7
Yi-34B-Chat	69.3/60.3	52.2/23.7	39.0/7.3	32.9/2.7
Deepseek-67B-Chat	78.1/72.6	63.8/33.0	53.5/19.0	60.9/12.7
Qwen-72B-Chat	89.4/71.0	76.9/52.7	64.9/30.7	66.3/15.3
△Mathematical Models				
MammoTH-7B	11.8/24.3	7.6/3.0	8.3/1.3	6.3/1.0
Metamath-Llemma-7B	21.2/49.3	23.3/9.0	22.0/9.0	16.1/4.0
InternLM2-Chat-Math-7B	64.9/67.0	57.7/40.3	50.7/18.0	46.5/7.3
Deepseek-Math-7B-Instruct	73.3/74.0	54.7/29.7	48.5/21.3	50.2/9.7
Deepseek-Math-7B-RL	78.9/82.7	69.7/44.7	59.9/31.0	68.0/17.3
MammoTH-13B	27.9/41.7	16.4/5.0	15.7/4.0	16.8/4.3
InternLM2-Chat-Math-20B	72.2/70.0	68.0/43.0	59.4/24.3	52.5/11.3
MammoTH-70B	57.9/60.7	45.2/11.0	38.9/8.3	43.7/5.3

C Extra Analysis

C.1 Prompts Demonstration

Please refer to the respective prompt block for a detailed demonstration.

C.1.1 English Open-ended test

The corresponding prompt is presented in Figure 17.

C.1.2 Chinese Open-ended test

The corresponding prompt is presented in Figure 18.

C.1.3 English single choice with reasoning

The corresponding prompt is presented in Figure 19.

C.1.4 Chinese single choice with reasoning

The corresponding prompt is presented in Figure 20.

C.2 Error Types Demonstration

Please refer to the respective cases for a detailed error types demonstration.

C.2.1 Misunderstandings of concepts

The corresponding case is presented in Figure 8.

C.2.2 Flawed reasoning

The corresponding case is presented in Figure 9.

C.2.3 Misaligned with the question

The corresponding case is presented in Figure 10.

C.2.4 Exceed max out length

The corresponding case is presented in Figure 11.

C.2.5 Responses constrained to Options

The corresponding case is presented in Figure 12.

C.2.6 Non-adherence to the prompt

The corresponding case is presented in Figure 13.

C.3 Reasoning Paths Demonstration

C.3.1 Small-scale chat model

The corresponding case is presented in Figure 14.

C.3.2 Large-scale chat model

The corresponding case is presented in Figure 15.

Table 6: Detailed Results of Bilingual Score on MathBench (EN/CN).

Models	Primary	Middle	High	College
★API Models				
GPT-3.5	76.5/59.3	39.2/25.3	41.7/33.4	42.9/21.5
GPT-4	79.1/86.6	67.3/52.3	66.1/67.2	63.4/54.7
♡Open-Source Chat Models				
ChatGLM3-6B	44.4/36.7	19.7/13.8	16.4/15.9	6.2/6.3
Yi-6B-Chat	42.9/39.2	19.5/17.5	13.0/13.7	9.6/8.5
InternLM2-Chat-7B	67.4/67.5	36.2/32.5	36.2/36.4	29.6/15.5
Qwen-7B-Chat	48.9/51.4	24.8/26.2	22.8/32.0	19.5/15.1
Deepseek-7B-Chat	43.4/35.2	16.3/11.7	12.9/9.2	9.7/8.7
Baichuan2-13B-Chat	54.9/53.6	25.3/23.0	18.5/25.4	17.6/11.1
Qwen-14B-Chat	64.8/67.4	36.9/42.6	33.1/51.4	27.6/26.8
InternLM2-Chat-20B	75.2/64.8	47.6/32.3	40.9/41.6	30.6/16.0
Yi-34B-Chat	62.0/67.6	28.6/31.6	26.3/35.7	16.5/19.1
Deepseek-67B-Chat	80.2/74.3	47.1/36.0	43.0/43.2	50.6/33.6
Qwen-72B-Chat	79.0/81.1	53.1/54.6	46.8/70.7	41.3/40.4
△Mathematical Models				
MammoTH-7B	26.8/9.4	8.2/1.1	8.5/2.5	6.7/0.7
Metamath-Llemma-7B	47.3/23.3	21.7/6.5	24.2/10.9	14.3/5.9
InternLM2-Chat-Math-7B	67.9/67.0	33.2/46.8	31.5/64.2	29.4/28.5
Deepseek-Math-7B-Instruct	71.5/75.8	33.6/39.1	32.8/48.7	27.8/32.0
Deepseek-Math-7B-RL	83.1/78.4	54.1/45.3	48.5/57.4	46.8/38.6
MammoTH-13B	49.6/19.9	17.7/3.4	15.1/5.0	15.9/5.2
InternLM2-Chat-Math-20B	69.4/69.9	36.4/47.2	33.7/68.3	30.8/29.1
MammoTH-70B	70.2/48.3	31.0/19.9	29.9/22.7	30.4/18.6

C.3.3 Math model

The corresponding case is presented in Figure 16.

C.3.1 Misunderstandings of Concepts

Prompt:

Question: Let R be a ring and let U and V be (two-sided) ideals of R . Which of the following must also be ideals of R ?

- I. $\{u + v : u \in U \text{ and } v \in V\}$
- II. $\{uv : u \in U \text{ and } v \in V\}$
- III. $\{x : x \in U \text{ and } x \in V\}$

Options:

- (A) II only ✗
- (B) III only ✗
- (C) I and II only ✗
- (D) I and III only ✓

Incorrect Model Answer:

1. $\{u + v : u \in U \text{ and } v \in V\}$ is not necessarily an ideal of R . This is because the set may not be closed under the ring operation. For example, if $u \in U$ and $v \in V$, but $u + v \notin U$ or $u + v \notin V$, then $u + v$ may not be in the set $\{u + v : u \in U \text{ and } v \in V\}$.

2. $\{uv : u \in U \text{ and } v \in V\}$ is an ideal of R . This is because for any $r \in R$, $u \in U$, and $v \in V$, we have $r(uv) = (ru)v \in UV$, and $(uv)r = u(vr) \in UV$. Thus, UV is closed under left and right multiplication by elements of R . Also, if $u_1v_1, u_2v_2 \in UV$, then $(u_1v_1) - (u_2v_2) = u_1v_1 - u_2v_2 = u_1v_1 + (-u_2v_2) \in UV$, since $-u_2v_2 \in UV$ because V is an ideal. Therefore, UV is an ideal of R .

3. $\{x : x \in U \text{ and } x \in V\}$ is not necessarily an ideal of R . This is because the set may not be closed under the ring operation. For example, if $u \in U$ and $v \in V$, but $uv \notin U$ or $uv \notin V$, then uv may not be in the set $\{x : x \in U \text{ and } x \in V\}$.

Therefore, the correct answer is option II. The answer is: A

Error Analysis:

In crafting its response, the model fails to accurately grasp the concept of an ideal within a ring.

Figure 8: An example of misunderstandings of mathematical concepts.

C.3.2 Flawed Reasoning

Prompt:

Question: Which of the following circles has the greatest number of points of intersection with the parabola $x^2 = y + 4$?

Options:

- (A) $x^2 + y^2 = 2$ ✗
- (B) $x^2 + y^2 = 9$ ✓
- (C) $x^2 + y^2 = 16$ ✗
- (D) $x^2 + y^2 = 1$ ✗

Incorrect Model Answer:

Step 1: Understand the problem We are given a parabola, $x^2 = y + 4$, and four circles. We need to find which circle has the maximum number of points of intersection with the parabola.

Step 2: Analyze the options Let's analyze each circle one by one:

Option A: $x^2 + y^2 = 2$ This equation represents a circle centered at the origin $(0, 0)$ with radius $\sqrt{2}$. To find the points of intersection between this circle and the parabola, we can substitute $y = x^2 - 4$ into the circle's equation and see if there are any real solutions for x .

Option B: $x^2 + y^2 = 9$ This equation represents a circle centered at the origin $(0, 0)$ with radius 3. We can follow the same process as in Option A to find the points of intersection.

Option C: $x^2 + y^2 = 16$ This equation represents a circle centered at the origin $(0, 0)$ with radius 4. Again, we can substitute $y = x^2 - 4$ into the circle's equation and look for real solutions.

Option D: $x^2 + y^2 = 1$ This equation represents a circle centered at the origin $(0, 0)$ with radius 1. We will follow the same process as before.

Step 3: Compare the results After analyzing all the options, we find that:

Option A has two points of intersection.

Option B has two points of intersection.

Option C has four points of intersection.

Option D has no points of intersection.

Step 4: Choose the best answer Based on our analysis, Option C has the greatest number of points of intersection with the parabola. Therefore, the correct answer is option C.

Error Analysis:

The model accurately grasps the mathematics of parabolas and circles, using equations to identify roots correctly. However, errors in subsequent reasoning led to inaccuracies in calculating the intersections between circles of various sizes and the specified parabola. Therefore, we classify this as a correct pathway of reasoning but with errors.

Figure 9: An example of flawed Reasoning despite logical consistency.

C.3.3 Misaligned with the Question

Prompt:

Question: For how many positive integers k does the ordinary decimal representation of the integer $k!$ end in exactly 99 zeros? A. None B. One C. Four D. Five

Options:

- (A) None
- (B) One
- (C) Four
- (D) Five

X
X
X
✓

Incorrect Model Answer:

Let the number of trailing zeros be m . The number is in the form of $10^m = (2^m)(5^m)$. m must be a multiple of both 2 and 5. There are four such numbers: 0, 10, 20, and 30. The answer is C.

Error Analysis:

The model overlooks discussing the tails of the factorial of decimal numbers in the problem and instead draws an irrelevant conclusion.

Figure 10: An example of reasoning misaligned with the query's intent.

C.3.4 Exceed max out length

Prompt:

Question: Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5 < Y_6$ be the order statistics of a random sample of size $n = 6$ from a distribution of the continuous type having $(100p)$ th percentile π_p . Compute

(a) $P(Y_2 < \pi_{0.5} < Y_5)$.

Options:

(A) -2.7812

X

(B) 0.7812

✓

(C) 0.7762

X

(D) -0.7028

X

Incorrect Model Answer:

1. The order statistics $Y_1 < Y_2 < Y_3 < Y_4 < Y_5 < Y_6$ of a random sample of size $n = 6$ from a distribution of the continuous type having $(100p)$ th percentile π_p .

2. We want to compute $P(Y_2 < \pi_{0.5} < Y_5)$.

3. The probability $P(Y_2 < \pi_{0.5} < Y_5)$ is equal to the probability that the 2nd order statistic is less than the 50th percentile (median) and the 5th order statistic is greater than the 50th percentile (median).

4. Using the concept of order statistics, we know that Y_2 is the 2nd smallest value in the sample, Y_5 is the 5th smallest value in the sample, and $\pi_{0.5}$ is the median of the distribution.

5. The probability $P(Y_2 < \pi_{0.5} < Y_5)$ is equal to the probability that exactly 2 out of the 6 sample values are less than the median and exactly 5 out of the 6 sample values are greater than the median.

6. This is a binomial probability problem with parameters $n = 6$, $p = 0.5$, and we want to find $P(X = 2)$ and $P(X = 5)$, where X is the number of sample values less than the median.

7. Using the binomial probability formula, $P(X = 2) = \binom{6}{2}(0.5)^2(0.5)^4 = 15 \times 0.25 \times 0.0625 = 0.234375$.

8. Similarly, $P(X = 5) = \binom{6}{5}(0.5)^5(0.5)^1 = 6 \times 0.03125 \times 0.5 = 0.09375$.

9. The probability $P(Y_2 < \pi_{0.5} < Y_5)$ is the product of these two probabilities,

Error Analysis:

The model demonstrates a clear grasp of quartiles and probability distributions. However, due to extensive reasoning rounds, responses often exceed the text length limit, resulting in truncation. Additionally, there is an error in the model's final statement concerning the probability product.

Figure 11: An example of responses exceeding the predefined character limit.

C.3.5 Responses Constrained to Options

Prompt:

Question: If $\mathbf{x} = \begin{pmatrix} 2 \\ 3i \\ 1 - i \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} -1 + i \\ 4 \\ -1 - i \end{pmatrix}$, find (\mathbf{y}, \mathbf{y}) .

Options:

- (A) 16.0
- (B) 16.7
- (C) -17.9
- (D) -18.6

✓

✗

✗

✗

Incorrect Model Answer:

To find the dot product of two vectors, we take the sum of the products of their corresponding components. Therefore, the correct answer is option B. 16.7

Error Analysis:

The model directly provides an answer without elaborating on the reasoning process behind it, making it challenging to ascertain the cause of the incorrect response.

Figure 12: An example of responses constrained to options.

C.3.6 Non-adherence to the Prompt

Prompt:

Question: Let M be a 5×5 real matrix. Exactly four of the following five conditions on M are equivalent to each other. Which of the five conditions is equivalent to NONE of the other four?

Options:

- (A) For any two distinct column vectors u and v of M , the set $\{u, v\}$ is linearly independent. ✓
- (B) The homogeneous system $Mx = 0$ has only the trivial solution. ✗
- (C) The system of equations $Mx = b$ has a unique solution for each real 5×1 column vector b . ✗
- (D) The determinant of M is nonzero. ✗

Incorrect Model Answer:

Let's analyze condition A. If for any two distinct column vectors u and v of M , the set $\{u, v\}$ is linearly independent, then it means that no column vector can be written as a linear combination of the others. This implies that the columns of M are linearly independent.

Step 2: Now let's consider condition B. The homogeneous system $Mx = 0$ has only the trivial solution if and only if the null space of M consists only of the zero vector. This is equivalent to saying that the columns of M are linearly independent.

Step 3: Next, we look at condition C. The system of equations $Mx = b$ has a unique solution for each real 5×1 column vector b if and only if the rank of M is equal to 5. Since M is a 5×5 matrix, this means that the columns of M must be linearly independent.

Step 4: We now examine condition D. The determinant of M is nonzero if and only if the columns of M are linearly independent.

Step 5: From steps 1-4, we see that conditions A, B, C, and D are all equivalent to each other, meaning they are all equivalent to the statement "the columns of M are linearly independent." Therefore, the condition that is equivalent to NONE of the other four is condition E, which is not given in the options.

Therefore, the correct answer is option E (not given).

Error Analysis:

The model does not provide answers in the given options.

Figure 13: An example of non-adherence to the prompt.

C.4.1 Reasoning Path of the Small-Scale Chat Model

Prompt:

Question: Alma bought a laptop computer at a store that gave a 20 percent discount off its original price. The total amount she paid to the cashier was p dollars, including an 8 percent sales tax on the discounted price. Which of the following represents the original price of the computer in terms of p ?

Options:

- (A) $\frac{p}{0.88}$ ✗
(B) $(0.8)(1.08)p$ ✗
(C) $\frac{p}{(0.8)(1.08)}$ ✓
(D) $0.88p$ ✗

Model Answer:

The original price of the laptop is x . After the 20% discount, the price becomes $0.8x$. We also know that the total amount Alma paid to the cashier, including the 8% sales tax, was p . So we can set up the equation:

$$0.8x + 0.08x = p$$

Simplifying this equation, we get:

$$0.96x = p$$

Now we can solve for x :

$$x = p/0.96$$

$$x = 0.88p$$

Therefore, the original price of the computer in terms of p is $0.88p$. Option D is the correct answer.

Reasoning Path Analysis:

The model's reasoning is predicated on an inaccurate interpretation of the tax calculation. Nevertheless, the mathematical reasoning employed by the model is comprehensive and logically coherent despite simple errors, indicating a sound mathematical logic within the reasoning trajectory of the small-scale chat model.

Figure 14: An example of reasoning path of the small-scale chat model.

C.4.2 Reasoning Path of the Large-Scale Chat Model

Prompt:

Question: It is known that $\log_2(x) = 5$. Give the value of $\log_2(2x)$. Explain your answer.

Options:

- (A) 6
- (B) 12
- (C) 36
- (D) 7

✓
✗
✗
✗

Model Answer:

Step 1: We are given that $\log_2(x) = 5$.

Step 2: Our goal is to find the value of $\log_2(2x)$.

Step 3: Let's use the property of logarithms that says $\log_b(mn) = \log_b(m) + \log_b(n)$.

Step 4: In our case, we can rewrite $2x$ as $2 \cdot x$. So, we have:

$$\log_2(2x) = \log_2(2) + \log_2(x)$$

Step 5: We know that $\log_2(2) = 1$, because $2^1 = 2$. And we also know that $\log_2(x) = 5$ (from the problem statement). So, we can substitute these values in:

$$\log_2(2x) = 1 + 5$$

Step 6: Adding 1 and 5 gives us:

$$\log_2(2x) = 6$$

Therefore, the correct answer is option A.

Reasoning Path Analysis:

The model's reasoning paths exhibit evident recursive relationships, cohesive logical connections, and well-defined objectives. Additionally, the model demonstrates adept utilization of pertinent knowledge references while responding to questions, thereby enhancing its mathematical comprehension and facilitating the provision of comprehensive and lucid reasoning paths.

Figure 15: An example of reasoning path of the large-scale chat model.

C.4.3 Reasoning Path of the Math Model

Prompt:

Question: The first term of an arithmetic sequence is 18. The sum of the first six terms of the sequence is equal to the sum of the first seven terms.

Show that the sum of the first 13 terms of the sequence is 0 and calculate the 13th term of the sequence.

Options:

(A) -18

✓

(B) -16

✗

(C) -14

✗

(D) -20

✗

Model Answer:

13th term of the sequence is 0.

Let a be the first term and d be the common difference.

The sum of the first 6 terms is:

$$S_6 = 6/2[2a + (6 - 1)d] = 3(2a + 5d) = 6a + 15d.$$

The sum of the first 7 terms is:

$$S_7 = 7/2[2a + (7 - 1)d] = 7/2(2a + 6d) = 7(a + 3d) = 7a + 21d.$$

Given that $S_6 = S_7$, we have:

$$6a + 15d = 7a + 21d.$$

Rearranging terms, we get:

$$a = -6d.$$

The sum of the first 13 terms is:

$$S_{13} = 13/2[2a + (13 - 1)d] = 13/2(2a + 12d) = 13(a + 6d).$$

Substituting $a = -6d$, we get:

$$S_{13} = 13(-6d + 6d) = 13 \times 0 = 0.$$

The n th term of an arithmetic sequence is given by:

$$a_n = a + (n - 1)d.$$

Substituting $a = -6d$, we get:

$$a_n = -6d + (n - 1)d = d(-6 + n - 1) = d(n - 7).$$

For the 13th term, we have:

$$a_{13} = d(13 - 7) = 6d.$$

But we know that $a = -6d$, so:

$$a_{13} = 6d = -a.$$

Since the first term is 18, we have $a = 18$, so:

$$a_{13} = -18.$$

Therefore, the correct answer is option A. The answer is .

Reasoning Path Analysis:

The mathematical model correctly invokes the formulas and properties of the equivarient series in its reasoning path and uses mathematical reasoning to draw conclusions. It correctly applied the summation and generalization formulas of the equal difference series and used algebraic operations to derive the results. The model was derived without logical errors, each step was based on the results of previous reasoning, and the derivation was consistent with the known conditions.

Figure 16: An example of reasoning path of the large-scale chat model.

C.1.1 English Open-ended test

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

BOT:'A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted $21 - 15 = 6$ trees. The answer is 6.'

HUMAN:'Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?'

BOT:'A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.'

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

BOT:"A: Leah had 32 chocolates and Leah's sister had 42. That means there were originally $32 + 42 = 74$ chocolates. 35 have been eaten. So in total they still have $74 - 35 = 39$ chocolates. The answer is 39."

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

BOT:'A: Jason had 20 lollipops. Since he only has 12 now, he must have given the rest to Denny. The number of lollipops he has given to Denny must have been $20 - 12 = 8$ lollipops. The answer is 8.'

HUMAN:'Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?'

BOT:'A: He has 5 toys. He got 2 from mom, so after that he has $5 + 2 = 7$ toys. Then he got 2 more from dad, so in total he has $7 + 2 = 9$ toys. The answer is 9.'

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

BOT:'A: There are 4 days from monday to thursday. 5 computers were added each day. That means in total $4 * 5 = 20$ computers were added. There were 9 computers in the beginning, so now there are $9 + 20 = 29$ computers. The answer is 29.'

HUMAN:'Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?'

BOT:'A: Michael initially had 58 balls. He lost 23 on Tuesday, so after that he has $58 - 23 = 35$ balls. On Wednesday he lost 2 more so now he has $35 - 2 = 33$ balls. The answer is 33.'

HUMAN:'Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?'

BOT:'A: She bought 5 bagels for \$3 each. This means she spent $5 * \$3 = \15 on the bagels. She had \$23 in beginning, so now she has $\$23 - \$15 = \$8$. The answer is 8.'

HUMAN:'Q: {question}'

BOT:'A: {answer}'

Figure 17: An example prompt of English Open-ended test.

C.1.2 Chinese Open-ended test

HUMAN: 'Q: 林中有15棵树。林务工人今天将在林中种植树木。完成后，将有21棵树。林务工人今天种植了多少棵树？'

BOT: 'A: 我们从15棵树开始。后来有21棵树。差值必定是他们种植的树木数量。所以，他们必须种植了 $21 - 15 = 6$ 棵树。答案是6'

HUMAN: 'Q: 如果停车场有3辆车，又有2辆车进来，停车场里有多少辆车？'

BOT: 'A: 停车场已经有3辆车。又进来了2辆车。现在有 $3 + 2 = 5$ 辆车。答案是5'

HUMAN: 'Q: 黎恩有32块巧克力，她的妹妹有42块。如果他们吃了35块，他们总共剩下多少块？'

BOT: 'A: 黎恩有32块巧克力，Leah的妹妹有42块。这意味着原本有 $32 + 42 = 74$ 块巧克力。被吃掉了35块。所以他们总共还剩下 $74 - 35 = 39$ 块巧克力。答案是39'

HUMAN: 'Q: 杰森有20个棒棒糖。他给丹妮一些棒棒糖。现在Jason只剩下12个棒棒糖。杰森给丹妮多少个棒棒糖？'

BOT: 'A: 杰森有20个棒棒糖。因为他现在只剩下12个，所以他必须把剩下的都给了丹妮。他给丹妮的棒棒糖数量必定是 $20 - 12 = 8$ 个。答案是8'

HUMAN: 'Q: 莎莎有五个玩具。在圣诞节，他从他的爸爸和妈妈那里各得到了两个玩具。现在他有多少个玩具？'

BOT: 'A: 她有5个玩具。他从妈妈那里得到了2个，所以之后他有 $5 + 2 = 7$ 个玩具。然后他从爸爸那里得到了2个，所以总共他有 $7 + 2 = 9$ 个玩具。答案是9'

HUMAN: 'Q: 服务器房间里有九台电脑。从周一到周四每天增加五台电脑。现在服务器房里有多少台电脑？'

BOT: 'A: 从周一到周四有4天。每天增加5台电脑。这意味着总共增加了 $4 * 5 = 20$ 台电脑。一开始有9台电脑，所以现在有 $9 + 20 = 29$ 台电脑。答案是29'

HUMAN: 'Q: 迈克尔有58个高尔夫球。星期二，他丢失了23个高尔夫球。星期三，他又丢失了2个。星期三结束时他还剩下多少个高尔夫球？'

BOT: 'A: 迈克尔一开始有58个球。星期二他丢失了23个，所以之后他还剩下 $58 - 23 = 35$ 个球。星期三他又丢失了2个，所以现在他还剩下 $35 - 2 = 33$ 个球。答案是33'

HUMAN: 'Q: 奥利弗有23美元。她用每个3美元的价格买了五个百吉饼。她还剩下多少钱？'

BOT: 'A: 她以每个3美元的价格买了5个百吉饼。这意味着她在百吉饼上花费了 $5 * 3 = 15$ 美元。她一开始有23美元，所以现在她还剩下 $23 - 15 = 8$ 美元。答案是8'

HUMAN: 'Q: {question}'

BOT: 'A: {answer}'

Figure 18: An example prompt of Chinese Open-ended test.

C.1.3 English single choice with reasoning

"Here is a multiple-choice question about mathematics. Please reason through it step by step, and at the end, provide your answer option with 'Therefore, the correct answer is option X', Where 'X' is the correct option you think from A, B, C, D. Here is the question you need to answer:

{question}

Let's think step by step: "

Figure 19: An example prompt of English single choice with reasoning.

C.1.4 Chinese single choice with reasoning

"以下是一道关于数学的单项选择题，请你一步一步推理，并在最后用“所以答案为选项X”给出答案，其中“X”为选项A, B, C, D中你认为正确的选项。下面是你要回答的问题

{question}

让我们一步一步思考： "

Figure 20: An example prompt of Chinese single choice with reasoning.