ConceptMath: A Bilingual Concept-wise Benchmark for Measuring Mathematical Reasoning of Large Language Models

Yanan Wu*[1], Jie Liu*[2,3], Xingyuan Bu*[1], Jiaheng Liu#[1], Zhanhui Zhou[3], Yuanxing Zhang[1], Chenchen Zhang[1], Zhiqi Bai[1], Haibin Chen[1], Tiezheng Ge[1], Wanli Ouyang[2,3], Wenbo Su[1], Bo Zheng[1]

[1]Alibaba Group [2]The Chinese University of Hong Kong

[3]Shanghai AI Laboratory

{lixing.wyn, ljh411989}@taobao.com

Abstract

This paper introduces ConceptMath, a bilingual (English and Chinese), fine-grained benchmark that evaluates concept-wise mathematical reasoning of Large Language Models (LLMs). Unlike traditional benchmarks that evaluate general mathematical reasoning with an average accuracy, ConceptMath systematically organizes math problems under a hierarchy of math concepts, so that mathematical reasoning can be evaluated at different granularity with concept-wise accuracies. Based on our ConcepthMath, we then evaluate a broad range of LLMs, and we observe existing LLMs, though achieving high average accuracies on traditional benchmarks, exhibit significant performance variations across different math concepts and may even fail catastrophically on the most basic ones. Besides, we also introduce an efficient fine-tuning strategy to enhance the weaknesses of existing LLMs. Finally, we hope ConceptMath could guide the developers to understand the fine-grained mathematical abilities of their models and facilitate the growth of foundation models. Code is available at https://github.com/conceptmath/conceptmath.

1 Introduction

Mathematical reasoning is a crucial capability for Large Language Models (LLMs). Recent advancements in LLMs, including Anthropic (Anthropic, 2023), GPT-4 (OpenAI, 2023), and LLaMA (Touvron et al., 2023a), have demonstrated impressive mathematical reasoning on existing benchmarks with high average accuracies on datasets like GSM8K (Cobbe et al., 2021). Although these benchmarks are able to measure the overall mathematical reasoning capabilities of LLMs *on average*, they fail to probe the fine-grained failure modes of mathematical reasoning *on specific*

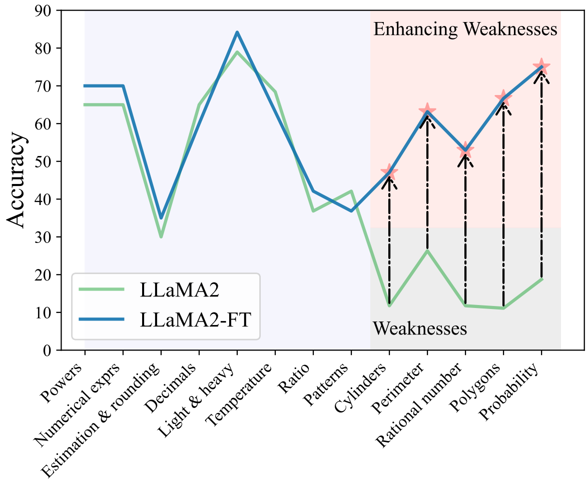


Figure 1: The concept-wise accuracies of LLaMA2-13B and the fine-tuned version based on our efficient fine-tuning method (i.e., LLaMA2-FT).

mathematical concepts. For example, Fig. 1 shows that the performance of LLaMA2-13B varies significantly across different concepts and fails on simple concepts like *Rational number* and *Cylinders*. It is crucial to know these specific failure modes of the language model, especially in some practical applications where we need to focus on specific mathematical abilities. For example, for financial analysts, calculation, measurement, and statistics are the concepts of most interest while others like geometry are not as important.

Moreover, the mathematics system, by its nature, is more fine-grained than holistic. It is typically organized into distinct math concepts [1], and humans develop comprehensive mathematical capabilities through a concept-by-concept, curriculum-based learning process (Simon, 2011; Fritz et al., 2013). These issues underscore the core motivation of this paper: *the need for a fine-grained benchmark that evaluates concept-wise mathematical reasoning of LLMs*.

Therefore, first, we introduce ConceptMath, the

* First three authors contributed equally. # Corresponding Author: Jiaheng Liu.

[1]https://en.wikipedia.org/wiki/Lists_of_mathematics_topics

Findings of the Association for Computational Linguistics: ACL 2024, pages 6815–6839

first bilingual (English and Chinese), concept-wise benchmark for measuring mathematical reasoning. ConceptMath gathers math concepts from four educational systems, resulting in four distinct mathematical concept systems: *English Elementary*, *English Middle*, *Chinese Elementary*, and *Chinese Middle* [2]. Each of these concept systems organizes around 50 atomic math concepts under a three-level hierarchy and each concept includes approximately 20 mathematical problems. Overall, ConceptMath comprises a total of 4011 math word problems across 214 math concepts, and Fig. 2 shows the diagram overview of ConceptMath.

Second, based on our ConceptMath, we perform extensive experiments to assess the mathematical reasoning of existing advanced LLMs, including 2 close-sourced LLMs and 17 open-sourced LLMs. These evaluations were performed in zero-shot, chain-of-thought (CoT), and few-shot settings. To our surprise, even though most of the evaluated LLMs claim to achieve high average accuracies on traditional mathematical benchmarks (e.g., GSM8K (Cobbe et al., 2021)), they fail catastrophically across a wide spectrum of mathematical concepts.

Third, to make targeted improvements on underperformed math concepts, we propose an efficient fine-tuning strategy by first training a concept classifier and then crawling a set of samples from a large open-sourced math dataset (Paster et al., 2023; Wang et al., 2023b) for further LLMs fine-tuning. In Fig. 1, for LLaMA2-FT, we observe that the results of these weaknesses improved a lot after using the efficient fine-tuning method.

In summary, our contributions are as follows:

- We introduce ConceptMath, the first bilingual, concept-wise benchmark for measuring mathematical reasoning. ConceptMath encompasses 4 systems, approximately 214 math concepts, and 4011 math word problems, which can guide further improvements on the mathematical reasoning of existing models.
- Based on ConceptMath, we evaluate many LLMs and perform a comprehensive analysis of their results. For example, we observe that most of these LLMs (including open-sourced, closed-sourced, general-purpose, or math-specialized models) show significant variations in their performance results across math concepts.
- We also evaluate the contamination rate of our ConceptMath and introduce a simple and efficient fine-tuning method to improve the weaknesses of existing LLMs.

[2] The four concept systems are abbreviated as **Elementary-EN**, **Middle-EN**, **Elementary-ZH**, and **Middle-ZH**.

2 ConceptMath

ConceptMath is the first bilingual, concept-wise benchmark for measuring mathematical reasoning. In this section, we describe the design principle, dataset collection process, dataset statistics and an efficient fine-tuning strategy to enhance the weaknesses identified by our ConceptMath.

2.1 Design Principle

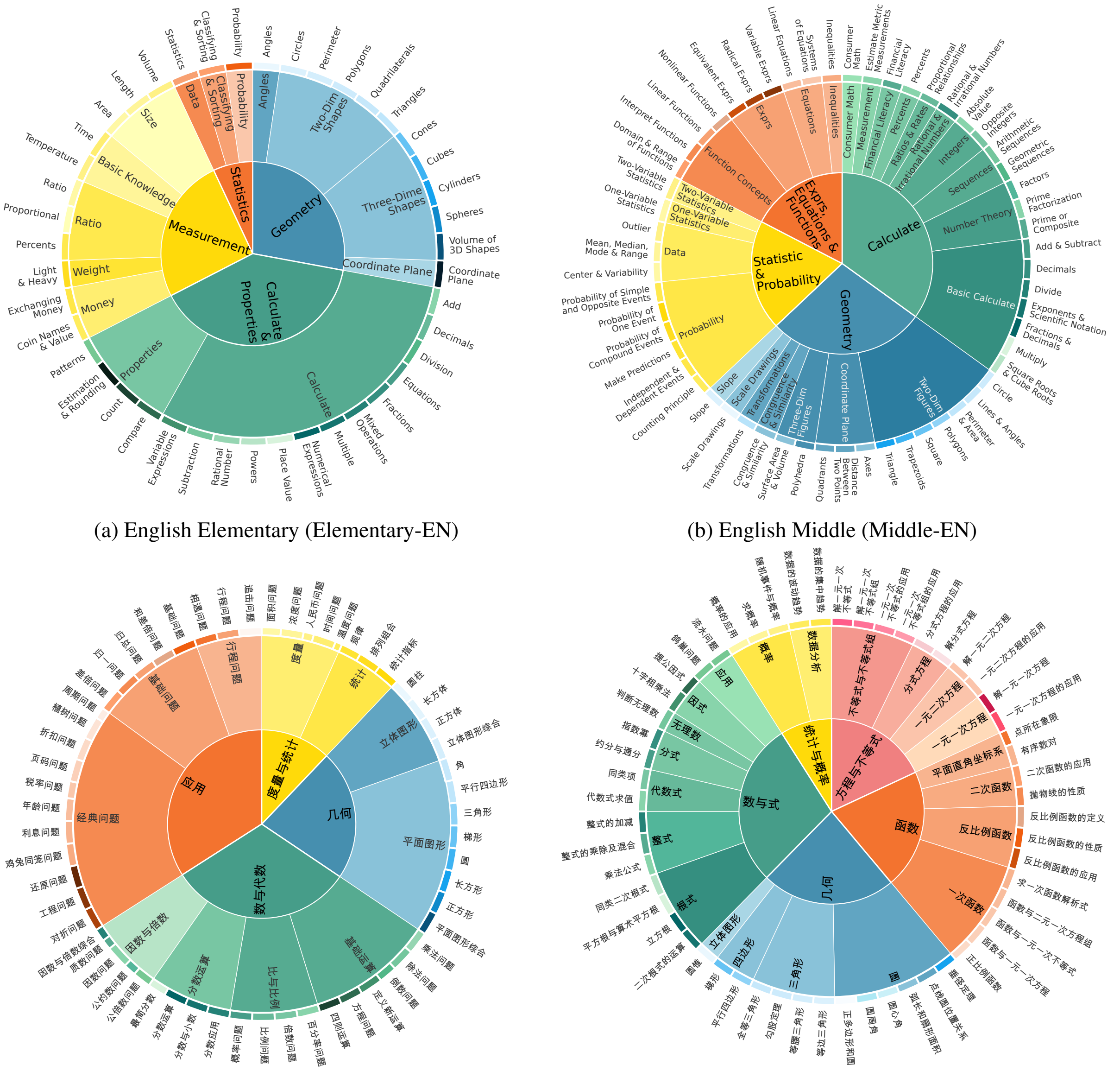
We created ConceptMath based on the following two high-level design principles:

Concept-wised Hierarchical System. The primary goal of ConceptMath is to evaluate the mathematical reasoning capacities of language models at different granularity. Therefore, ConceptMath organizes math problems within a three-level hierarchy of mathematical concepts in Fig. 2. This approach provides concept-wise evaluation for mathematical reasoning of language models and makes targeted and effective improvements possible.

Bilingualism. Most of the current mathematical benchmark focuses solely on English, leaving multi-lingual mathematical reasoning unexplored. As an early effort to explore multi-lingual mathematical reasoning, we evaluate mathematical reasoning in two languages: English and Chinese. Besides, since cultures and educational systems vary across different languages, common math concepts can differ a lot. Therefore, we carefully collect concepts in both languages, instead of merely translating from one language to another. For example, measurement metrics (e.g., money, size) are different for English and Chinese.

2.2 Data Collection

Subsequently, for data collection, we take a two-step approach to operationalize the aforementioned design principles: First, we recruit experts to delineate a hierarchy of math concepts based on different education systems. Secondly, we collect problems for each concept from various sources or design problems manually, which is succeeded by quality assessment and data cleaning.



(a) English Elementary (Elementary-EN)

(b) English Middle (Middle-EN)

(c) Chinese Elementary (Elementary-ZH)

(d) Chinese Middle (Middle-ZH)

Figure 2: Diagram overview of four concept systems in ConceptMath. We have provided translated Chinese concept names in English (See Appendix D).

Math Concept System Construction. Since the education systems provide a natural hierarchy of math concepts, we recruited four teachers from elementary and middle schools, specializing in either English or Chinese, to organize a hierarchy of math concepts for different education systems. This leads to four concept systems: Elementary-EN, Middle-EN, Elementary-ZH, and Middle-ZH, with each system consisting of a three-level hierarchy of around 50 atomic math concepts (Fig. 2).

Math Problem Construction. Then, we conducted a thorough data acquisition from various sources (including educational websites, textbooks, and search engines with specific concepts) to collect math word problems (including both questions and answers) for each math concept. To guarantee a balance across all concepts, approximately 20 problems were gathered for each math concept. Following this, both GPT-4 (OpenAI, 2023) and human experts were employed to verify and rectify the categorization and the solution of each problem. However, we observed that for some concepts, the problem count was significantly below 20. To address this long-tail issue, manual efforts were undertaken to augment these categories, ensuring a consistent collection of 20 problems for each concept. Furthermore, to broaden the diversity of the dataset and minimize the risk of data contamination, all gathered problems were paraphrased using GPT-4. It is impor-

Benchmark	Language	Fine-grained	Size
GSM8K (Cobbe et al., 2021)	EN	✗	1319
MATH (Hendrycks et al., 2021c)	EN	✗	5000
TabMWP (Lu et al., 2023)	EN	✗	7686
Dolphin18K (Huang et al., 2016)	EN	✗	1504
Math23K (Wang et al., 2017)	ZH	✗	1000
ASDiv (Miao et al., 2020)	EN	✗	2305
SingleOp (Roy et al., 2015)	EN	✗	562
AddSub (Hosseini et al., 2014)	EN	✗	395
MultiArith (Roy and Roth, 2015)	EN	✗	600
MMLU-Math (Hendrycks et al., 2021a)	EN	✗	906
Bilingual MWPs (Tan et al., 2021)	EN&ZH	✗	1557
ConceptMath	EN&ZH	✓	4011

Table 1: A comparison of our ConceptMath with some notable mathematical datasets. Note that the size is the number of samples of the test split.

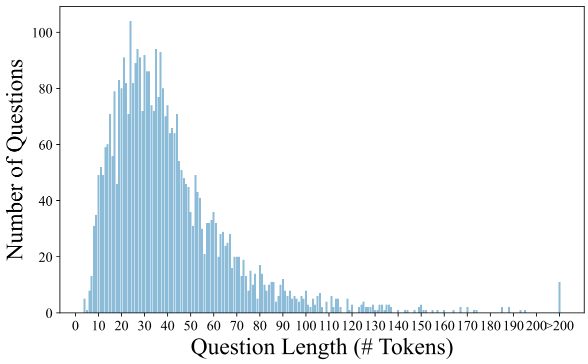


Figure 3: Length distributions of our ConceptMath.

tant to note that the collection and annotation processes were carried out by a team of six members, each possessing a university degree in an engineering discipline, to maintain a high level of technical expertise in executing these tasks.

2.3 Dataset Statistics

Comparison to existing datasets. As shown in Table 1, our ConceptMath differs from related datasets in various aspects: (1) ConceptMath is the first dataset to study fine-grained mathematical concepts and encompasses 4 systems, 214 math concepts, and 4011 math word problems. (2) Problems in ConcepthMath are carefully annotated based on the mainstream education systems for English (EN) and Chinese (ZH).

Details on the hierarchical system. Apart from Fig. 2, we also provide the details on the hierarchical system more clearly in Appendix D.

Length distribution. Fig. 3 shows the length distribution of our ConcepthMath, where number of tokens is reported [3]. The minimum, average and maximum of the tokens for these questions are 4, 41 and 309, respectively, which shows that they have lexical richness.

[3]We use the "cl100k_base" tokenizer from `https://github.com/openai/tiktoken`

2.4 Efficient Fine-Tuning

Based on our ConceptMath, we are able to identify the weaknesses in the mathematical reasoning capability of LLMs through concept-wise evaluation. In this section, we explore a straightforward approach to enhance mathematical abilities towards specific concepts by first training a concept classifier and then curating a set of samples from a large open-sourced math dataset. Specifically, first, by additionally collecting extra 10 problems per concept, we construct a classifier capable of identifying the concept class of a given question. The backbone of this classifier is a pretrained bilingual LLM, where the classification head is operated on its last hidden output feature. Then, we proceed to fine-tune LLMs using this specific dataset combined with the existing general math dataset, which aims to avoid overfitting on a relatively small dataset.

3 Experiments

In this section, we perform extensive experiments to demonstrate the effect of our ConceptMath.

3.1 Experimental Setup

Evaluated Models. We assess the mathematical reasoning of existing advanced LLMs on ConceptMath, including 2 close-sourced LLMs (i.e., GPT-3.5/GPT-4 (OpenAI, 2023)) and 17 open-sourced LLMs (i.e., WizardMath-13B (Luo et al., 2023), MetaMath-13B (Yu et al., 2023), MAmmoTH-13B (Yue et al., 2023), Qwen-14B/72B (Bai et al., 2023b), Baichuan2-13B (Baichuan, 2023), ChatGLM3-6B (Du et al., 2022), InternLM2-7B/20B (Team, 2023a), InternLM2-Math-7B/20B (Team, 2023a), LLaMA2-7B/13B/70B (Touvron et al., 2023b), Yi-6B/34B (Team, 2023b) and DeepSeekMath-7B (Shao et al., 2024)). Note that WizardMath-13B, MetaMath-13B, and MAmmoTH-13B are specialized math language models fine-tuned from LLaMA2. InternLM2-Math and DeepSeekMath are specialized math language models fine-tuned from corresponding language models. More details of these models can be seen in Appendix B.

Evaluation Settings. We employ three distinct evaluation settings: zero-shot, zero-shot with chain-of-thought (CoT), and few-shot promptings. The zero-shot prompting assesses the models' intrinsic problem-solving abilities without any prior examples. The zero-shot with CoT prompting

Model	Elementary-EN			Middle-EN			Elementary-ZH			Middle-ZH			Avg.
	ZS	ZS-COT	FS	ZS	ZS-COT	FS	ZS	ZS-COT	FS	ZS	ZS-COT	FS	
Yi-6B	67.94	67.56	59.03	65.55	64.59	56.05	34.33	31.91	37.86	36.46	36.19	36.46	49.49
ChatGLM3-6B	60.69	63.10	53.18	51.25	60.17	51.34	46.23	43.63	40.74	44.77	43.32	40.43	49.90
DeepSeekMath-7B	66.92	77.35	73.92	56.53	69.87	66.31	60.47	62.33	64.19	56.50	56.95	56.86	64.02
InternLM2-Math-7B	71.12	72.01	69.59	63.44	62.96	63.05	57.30	58.23	58.60	53.79	53.16	53.88	61.43
InternLM2-7B	68.83	69.97	66.67	37.04	65.83	55.47	47.63	49.02	53.02	45.22	45.40	44.86	54.08
LLaMA2-7B	36.51	42.62	38.68	34.26	39.16	33.69	15.72	17.67	17.58	30.87	32.22	27.80	30.57
MAmmoTH-13B	61.32	52.42	56.49	53.93	45.20	48.08	22.33	33.30	23.81	27.98	43.05	29.15	41.42
WizardMath-13B	41.73	44.78	34.99	36.85	37.72	45.11	10.51	11.26	18.70	12.36	15.52	22.92	27.70
MetaMath-13B	54.45	51.78	47.96	44.24	43.47	47.50	11.44	17.30	27.53	21.21	26.08	29.60	35.21
Baichuan2-13B	68.83	68.58	54.07	67.66	69.67	40.40	57.02	58.23	22.05	55.05	55.32	26.90	53.65
LLaMA2-13B	44.02	49.75	47.07	44.72	46.45	43.09	20.19	24.19	22.14	33.30	35.38	26.17	36.37
Qwen-14B	46.95	65.78	72.65	38.48	59.60	67.85	28.09	65.12	64.47	22.92	58.30	62.09	54.36
InternLM2-Math-20B	74.05	75.32	73.41	64.11	71.21	70.83	62.98	61.95	61.77	55.14	55.78	56.86	65.28
InternLM2-20B	53.31	72.52	73.28	45.11	67.47	56.72	48.19	55.53	59.81	45.13	50.63	56.68	57.03
Yi-34B	74.68	73.66	56.36	72.26	74.66	65.83	50.05	51.16	38.79	45.40	43.95	40.97	57.31
LLaMA2-70B	56.11	60.31	30.53	58.06	60.94	31.67	28.65	26.70	24.37	37.64	34.30	28.43	39.81
Qwen-72B	77.10	75.06	77.23	74.66	69.87	73.99	71.16	68.65	61.86	71.30	65.43	62.45	70.73
GPT-3.5	85.75	92.37	84.35	83.88	90.12	82.73	56.47	53.21	56.93	51.90	53.52	55.69	70.58
GPT-4	86.77	90.20	89.57	84.26	89.83	88.68	67.91	72.28	72.00	63.81	64.26	66.61	78.02
Avg.	63.00	66.59	61.00	56.65	62.57	57.28	41.93	45.35	43.49	42.67	45.72	43.41	52.47

Table 2: Results of different models on our constructed ConceptMath benchmark dataset. Note that "ZS", "ZS-COT", "FS" represents "zero-shot", "zero-shot w/ chain-of-thought" and "few-shot", repsectively. Models are grouped roughly according to their model sizes.

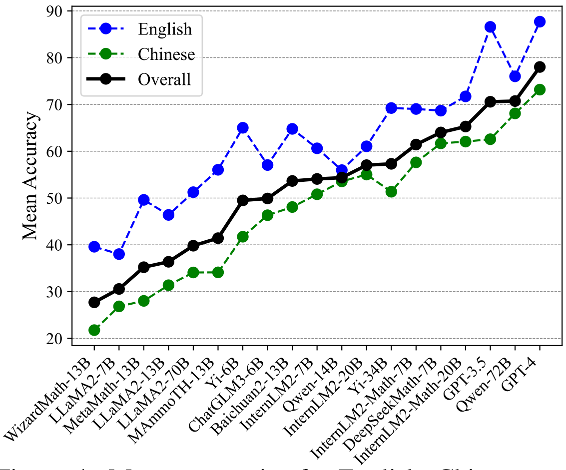


Figure 4: Mean accuracies for English, Chinese, and overall educational systems.

evaluates the models' ability to employ a logical chain of thought. In the few-shot prompting setting, the model is provided with fixed 5-shot prompts for different systems (See Appendix E), which includes five newly created examples with concise ground truth targets. This approach is designed to measure the in-context learning abilities. Besides, following MATH (Hendrycks et al., 2021c), all questions and answers in ConceptMath have been carefully curated, and each problem is evaluated based on exact matches. Moreover, greedy decoding with a temperature of 0 is used.

3.2 Results

Overall Accuracy. We present the overall accuracies of different LLMs on our ConceptMath benchmark under various prompt settings in Table 2. Subsequently, we analyzed the mathematical abilities of these LLMs in both English and Chinese in Fig. 4. Our analysis led to the following key findings: (1) GPT-3.5/4 showcases the most advanced mathematical reasoning abilities among LLMs in both English and Chinese systems, and the leading open-source Qwen-72B model archives comparable performance compared with GPT-3.5. (2) The scores on Chinese systems of most existing LLMs are lower than English systems a lot. For example, accuracies on Middle-ZH and Middle-EN for GPT-4 are 63.81 and 84.26. (3) Several models (e.g., WizardMath-13B or MetaMath-13B) fine-tuned from LLaMA2-13B achieve slight improvements on English systems, but the performance results are lower than LLaMA2-13B on Chinese systems a lot, which indicates that domain-specific fine-tuning may degrade the generalization abilities of LLMs. (4). The mathematical models (i.e., InternLM2-Math-7B/20B and DeepSeekMath-7B) by continuing pretraining on the large-scale math-related dataset (>=100B tokens) show sufficient improvements when compared to models with similar size, which indicates that large-scale pertaining is effective to improve the mathematical reasoning abilities.

Mean Concept-wised Accuracy. We show the mean accuracy for each concept on the tested mod-

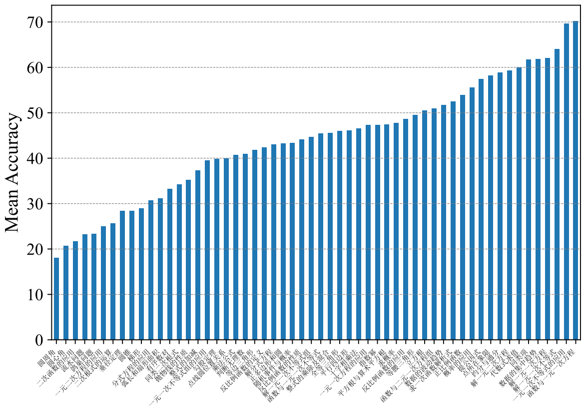


Figure 5: Mean concept accuracies on Middle-ZH.

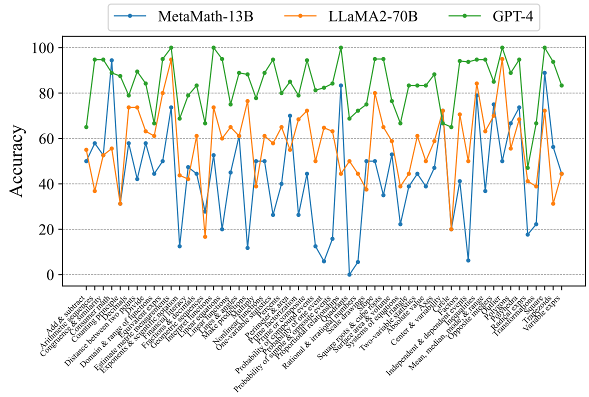


Figure 6: Concept accuracies on Middle-EN.

els using zero-shot prompting on Middle-ZH in Fig. 5 (See Appendix A for results on other concept systems). It shows that the accuracies across concepts vary a lot for existing LLMs. For example, around 18% of concepts of Middle-ZH exhibit an accuracy lower than 30%. These concepts with significant potential for improvement should be prioritized to effectively enhance the mathematical abilities of LLMs.

Concept-wised Accuracy. Fig. 6 shows that most existing LLMs, whether open-sourced, closed-sourced, general-purpose, or math-specialized, exhibit notable differences in their concept accuracies in the zero-shot prompt setting. These disparities may stem from variations in training datasets, strategies, and model sizes, which suggests that apart from common weaknesses, each model possesses its unique areas of deficiency or shortcomings. The concept accuracies of other concept systems and results of all models can be found in the Appendix A.

3.3 Analysis

Contamination. To determine whether a text is in the pretraining data of a LLM, we provide two different contamination detection methods (i.e., Rouge-based and Prob-based methods) to analyze our ConceptMath in Table 3. Specifically, for the Rouge-based method, we just input the first 50% of the question as the input and compute the Rouge-L score between the generation results and the ground-truth label of the last 50% of the text, where a lower Rouge-L score means a lower contamination rate. For the Prob-based method, we follow (Shi et al., 2023) to use the MIN-K% probability metric, which first gets the probability for each token in the test, and selects the K% tokens with minimum probabilities and calculate their average log-likelihood. If the average log-likelihood is high, the text is likely in the pretraining data. Note that we choose K as 10 in our setting. In Table 3, we observe that the contaminate rates on our ConceptMath are very low, which means that our ConceptMath can provide a reasonable evaluation for existing LLMs.

Unmastered Concepts. We also highlight the several unmastered concepts of the LLaMA2-13B in Table 4, which shows ConceptMath is effective in guiding further refinement of existing LLMs.

Evaluation Prompting. Different from the few-shot or cot prompting evaluation that can boost closed-source models, we find that zero-shot prompting is more effective for certain open-source LLMs in Table 2. This disparity may arise either because the models are not sufficiently powerful to own mathematical CoT capabilities (Yu et al., 2023; Wei et al., 2022) or because these models have already incorporated CoT data during training (Longpre et al., 2023). Consequently, to ensure a comprehensive analysis, we have employed all three prompting methods for evaluation.

Efficient Fine-tuning. To show the effect of efficient fine-tuning, we take the LLaMA2-13B as an example in Table 5. Specifically, we first select 10 concepts with the lowest accuracies in Elementary-EN. Then, we crawl 495 samples (about 50 samples per concept) using the trained classifier as the Concept-Specific (**CS**) training data from OpenWebMath (Paster et al., 2023). Meanwhile, to avoid overfitting, we introduce the MetaMathQA (**MMQA** (Yu et al., 2023)) data to preserve general mathematical abilities. After that, we can fine-tune LLaMA2-13B by only using MMQA (i.e., LLaMA2 (w/ MMQA)), or using both MMQA and CS data (i.e., LLaMA2 (w/ MMQA & CS)). In

Model	Elementary-EN	Middle-EN	Elementary-ZH	Middle-ZH	Avg. ↓
Yi-6B	5.30 / 1.73	5.21 / 1.37	0.04 / 0.20	0.36 / 0.35	2.73 / 0.91
ChatGLM3-6B	7.42 / 0.22	7.55 / 0.23	0.11 / 0.02	0.35 / 0.05	3.86 / 0.13
InternLM2-Math-7B	7.42 / 0.22	7.55 / 0.23	0.11 / 0.02	0.35 / 0.05	3.86 / 0.13
InternLM2-7B	5.36 / 1.03	5.27 / 0.84	0.01 / 0.37	0.33 / 0.49	2.74 / 0.68
MAmmoTH-13B	7.67 / 0.47	7.97 / 0.46	0.00 / 0.03	0.35 / 0.03	4.00 / 0.25
WizardMath-13B	8.41 / 0.35	8.23 / 0.34	0.00 / 0.02	0.55 / 0.02	4.30 / 0.18
MetaMath-13B	7.67 / 0.47	7.97 / 0.46	0.00 / 0.03	0.35 / 0.03	4.00 / 0.25
Baichuan2-13B	7.20 / 1.43	6.58 / 1.18	0.05 / 0.54	0.41 / 0.65	3.56 / 0.95
LLaMA2-13B	6.80 / 0.73	6.36 / 0.64	0.01 / 0.15	0.56 / 0.16	3.43 / 0.42
Qwen-14B	11.04 / 1.58	9.73 / 1.08	1.43 / 1.27	0.70 / 0.93	5.73 / 1.22
InternLM2-Math-20B	5.58 / 1.30	5.51 / 0.99	0.03 / 0.47	0.34 / 0.47	2.86 / 0.81
InternLM2-20B	7.20 / 1.43	6.58 / 1.18	0.05 / 0.54	0.41 / 0.65	3.56 / 0.95
GPT-3.5	9.48 / -	9.21 / -	0.00 / -	0.31 / -	4.75 / -
GPT-4	8.68 / -	8.24 / -	0.15 / -	0.68 / -	4.44 / -

Table 3: Data contamination rate of LLMs. We provide two different contamination detection methods. The values in the table represent "Rouge / Prob". Note that the second method based on output probability distributions can only be applied to the open-source models.

Concept	Concept Acc	Error Type	Example
Perimeter	26.32	Calculation Error	Q: Find the perimeter of a 16 by 18 feet rectangle. A: The perimeter is 16 + 18 + 16 + 18 = 72 feet.
		Knowledge Error	Q: The perimeter of a rectangular sticker is 14 centimeters. The sticker is 2 centimeters wide. How tall is it? A: The height equals the perimeter subtracts the width. The height of the rectangle is $14 - 12 = 12$ centimeters.
Polygons	11.11	Knowledge Error	Q: A regular polygon has an interior angle that measures 120 degrees. How many sides does this polygon have? A: Let n be the number of sides of the polygon. We have that $n \times 120 = 360$. Hence: $n = 3$.
Cylinders	11.76	Calculation Error	Q: How much soup can a cylinder-shaped can store, if the radius of the can is 3 inches and the height is 5 inches? A: ... The volume of the can is V = $\pi \times 3^2 \times 5 = 30\pi$.

Table 4: Error reasons of several unmastered concepts of LLaMA2-13B.

Table 5, we observe that LLaMA2 (w/ MMQA & CS) archives significant improvements on the lowest 10 concepts and preserves well on the other 33 concepts, which shows the effect of efficient fine-tuning and the advantages of our ConceptMath.

Classifying questions by concept. A key consideration for ConceptMath is whether it is appropriate to arrange evaluations by concept in LLMs because it is clear to humans that a question may involve one or several concepts. We verify if LLMs have this same level of understanding by computing the Mean Reciprocal Rank (MRR) [4] and accuracy of LLMs in classifying each question by concept. We use secondary categories of our hierarchical system as an example to show the MRR and accuracy of each knowledge system. There are approximately 18 concepts in the secondary categories for each knowledge system. As shown in Table 6, open-source and proprietary models can understand concepts to some degree, and the ability to understand concepts is positively correlated with the model's mathematical capabilities. The calculation method and metrics for each concept system are detailed in the Appendix C.

[4] https://en.wikipedia.org/wiki/Mean_reciprocal_rank

4 Related Work

Large Language Models for Mathematics. Large Language Models (LLMs) such as GPT-3.5 and GPT-4 have exhibited promising capabilities in complex tasks (Du et al., 2024; Guo et al., 2023b; Liu et al., 2024a; Zhang et al., 2024; Guo et al., 2024b; Sun et al., 2024). However, the proficiency of open-source alternatives like LLaMA (Touvron et al., 2023a) and LLaMA2 (Touvron et al., 2023b) remains notably inferior on these datasets, particularly in handling

Model	LLaMA2	LLaMA2 (w/ MMQA)	LLaMA2 (w/ MMQA &CS)
Cones	0.00	17.65	23.53
Spheres	5.88	29.41	35.29
Polygons	11.11	61.11	66.67
Rational Number	11.76	23.53	52.94
Cylinders	11.76	35.29	47.06
Angles	11.76	47.06	58.82
Probability	18.75	25.00	75.00
Perimeter	26.32	42.11	63.16
Volume	27.78	38.89	66.67
Proportional	27.78	33.33	44.44
Avg Acc. (over 10 concepts)	15.29	36.88	**53.36**
Avg Acc. (over 33 concepts)	51.94	58.14	**60.67**
Overall Acc.	44.02	53.94	**59.29**

Table 5: Results of fine-tuning models. MMQA and CS denote MetaMathQA and our constructed Concept-Specific training datasets, respectively. Introducing CS data specifically for the bottom 10 concepts significantly enhances these concepts' performance, while slightly improving the performance across the remaining 33 concepts.

Model	MRR	Accuracy
LLaMA2-7B	32.91	18.41
LLaMA2-13B	44.35	28.63
DeepSeekMath-7B	57.73	43.52
Qwen1.5-7B	56.58	42.32
Qwen1.5-14B	64.69	51.58
GPT-3.5	69.11	56.86
GPT-4	78.17	68.36

Table 6: The average MRR and accuracy of LLMs for classifying each question by concept.

non-English problems. In contrast, models like Baichuan2 (Baichuan, 2023) and Qwen (Bai et al., 2023b) pretrained on multilingual datasets (i.e., Chinese and English) have achieved remarkable performance. Recently, many domain-specialized math language models have been proposed. For example, MetaMath (Yu et al., 2023) leverages the LLaMA2 models and finetunes on the constructed MetaMathQA dataset. MAmmoTH (Yue et al., 2023) synergizes Chain-of-Thought (CoT) and Program-of-Thought (PoT) rationales.

Mathmatical Reasoning Benchmarks. Recently, many mathematical datasets (Roy and Roth, 2015; Koncel-Kedziorski et al., 2015; Lu et al., 2023; Huang et al., 2016; Miao et al., 2020; Patel et al., 2021) have been proposed. For example, SingleOp (Roy et al., 2015), expands the scope to include more complex operations like multiplication and division. GSM8K (Cobbe et al., 2021) is a widely used dataset, which requires a sequence of elementary calculations with basic arithmetic operations.

Fine-Grained Benchmarks. Traditional benchmarks focus on assessing certain abilities of models on one task (Guo et al., 2023b; Wang et al., 2023a; Liu et al., 2020; Guo et al., 2022; Chai et al., 2024; Liu et al., 2024a; Guo et al., 2024a, 2023c; Bai et al., 2023a; Liu et al., 2022; Guo et al., 2023a; Bai et al., 2024; Li et al., 2024). For example, the GLUE benchmark (Wang et al., 2019) combines a collection of tasks, and has witnessed superhuman model performance for pretraining models (Radford et al., 2019). Hendrycks et al. (2021b) introduced MMLU, a benchmark with multiple-choice questions across 57 subjects including STEM, humanities, and social sciences, for assessing performance and identifying weaknesses. Srivastava et al. (2022) proposed BIG-bench with over 200 tasks.

5 Conclusion

We introduce a new bilingual concept-wise math reasoning dataset called ConceptMath to assess models across a diverse set of concepts. First, ConceptMath covers more than 200 concepts across elementary and middle schools for mainstream English and Chinese systems. Second, we extensively evaluate existing LLMs by three prompting methods, which can guide further improvements for these LLMs on mathematical abilities. Third, we analyze the contamination rates, error cases and provide a simple and efficient fine-tuning strategy to enhance the weaknesses.

Limitations

Mathematical concepts may form a network, where failure to understand a complex concept can result from deficiencies in simpler foundational concepts or issues within the complex concept itself. Thus, the current hierarchical arrangement may be insufficient, and introducing an interrelatedness network of concepts could provide more insightful explanations of the mathematical abilities of language models. Building an accurate and interpretable interrelatedness network within our hierarchical system is a promising direction for future research.

Ethics Statement

ConceptMath consists of math problems, thus very few samples contain offensive content or sensitive personal information. All annotators were fully informed about the entire annotation process.

References

Anthropic. 2023. Model card and evaluations for claude models.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. arXiv preprint arXiv:2402.14762.

Jiaqi Bai, Hongcheng Guo, Jiaheng Liu, Jian Yang, Xinnian Liang, Zhao Yan, and Zhoujun Li. 2023a. Griprank: Bridging the gap between retrieval and generation via the generative knowledge improved passage ranking. CIKM.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023b. Qwen technical report. arXiv preprint arXiv:2309.16609.

Baichuan. 2023. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. arXiv preprint arXiv:2401.07037.

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

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Binhang Yuan, Wenhu Chen, Jie Fu, and Ge Zhang. 2024. Chinese tiny llm: Pretraining a chinese-centric large language model. Preprint, arXiv:2404.04167.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 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), pages 320–335.

Annemarie Fritz, Antje Ehlert, and Lars Balzer. 2013. Development of mathematical concepts as basis for an elaborated mathematical understanding. South African Journal of Childhood Education, 3(1):38–67.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, Zheng Cui, and Furu Wei. 2022. Lvp-m3: language-aware visual prompt for multilingual multimodal machine translation. EMNLP.

Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. 2023a. M2c: Towards automatic multimodal manga complement. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9876–9882.

Hongcheng Guo, Jian Yang, Jiaheng Liu, Jiaqi Bai, Boyang Wang, Zhoujun Li, Tieqiao Zheng, Bo Zhang, Qi Tian, et al. 2024a. Logformer: A pre-train and tuning pipeline for log anomaly detection. AAAI.

Hongcheng Guo, Jian Yang, Jiaheng Liu, Liqun Yang, Linzheng Chai, Jiaqi Bai, Junran Peng, Xiaorong Hu, Chao Chen, Dongfeng Zhang, et al. 2023b. Owl: A large language model for it operations. arXiv preprint arXiv:2309.09298.

Jinyang Guo, Jiaheng Liu, Zining Wang, Yuqing Ma, Ruihao Gong, Ke Xu, and Xianglong Liu. 2023c. Adaptive contrastive knowledge distillation for bert compression. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8941–8953.

Jinyang Guo, Jianyu Wu, Zining Wang, Jiaheng Liu, Ge Yang, Yifu Ding, Ruihao Gong, Haotong Qin, and Xianglong Liu. 2024b. Compressing large language models by joint sparsification and quantization. ICML.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 887–896.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. Transactions of the Association for Computational Linguistics, 3:585–597.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Zekun Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, et al. 2024. Cif-bench: A chinese instruction-following benchmark for evaluating the generalizability of large language models. arXiv preprint arXiv:2402.13109.

Jiaheng Liu, Zhiqi Bai, Yuanxing Zhang, Chenchen Zhang, Yu Zhang, Ge Zhang, Jiakai Wang, Haoran Que, Yukang Chen, Wenbo Su, et al. 2024a. E2-llm: Efficient and extreme length extension of large language models. arXiv preprint arXiv:2401.06951.

Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun, and Ping Li. 2022. Cross-lingual cross-modal consolidation for effective multilingual video corpus moment retrieval. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1854–1862.

Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Ken Chen, Wanli Ouyang, and Dong Xu. 2020. Block proposal neural architecture search. IEEE Transactions on Image Processing, 30:15–25.

Jie Liu, Zhanhui Zhou, Chao Yang, Han-Sen Zhong, and Wanli Ouyang. 2024b. Storm-7b: An empirical study of iterative direct preference optimization.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In The Eleventh International Conference on Learning Representations.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:2308.09583.

Norman Megill and David A Wheeler. 2019. Metamath: a computer language for mathematical proofs. Lulu. com.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 975–984.

OpenAI. 2023. Gpt-4 technical report. PREPRINT.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text. Preprint, arXiv:2310.06786.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1743–1752.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. Transactions of the Association for Computational Linguistics, 3:1–13.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. arXiv preprint arXiv:2309.15025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. arXiv preprint arXiv:2310.16789.

Martin A Simon. 2011. Studying mathematics conceptual learning: Student learning through their mathematical activity. North American Chapter of the International Group for the Psychology of Mathematics Education.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Tao Sun, Linzheng Chai, Yuwei Yin Jian Yang, Hongcheng Guo, Jiaheng Liu, Bing Wang, Liqun Yang, and Zhoujun Li. 2024. Unicoder: Scaling code large language model via universal code. ACL.

Minghuan Tan, Lei Wang, Lingxiao Jiang, and Jing Jiang. 2021. Investigating math word problems using pretrained multilingual language models. arXiv preprint arXiv:2105.08928.

InternLM Team. 2023a. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM-techreport`.

Yi Team. 2023b. Yi: Building the next generation of open-source and bilingual llms. `https://github.com/01-ai/Yi`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In Proceedings of the 2017 conference on empirical methods in natural language processing, pages 845–854.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv: 2310.00746.

Zengzhi Wang, Rui Xia, and Liu Pengfei. 2023b. Generative ai for math: Part i – mathpile: A billion-token-scale pretraining corpus for math. arXiv preprint arXiv:2312.17120.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

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

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. arXiv preprint arXiv: 2405.19327.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. arXiv preprint arXiv:2310.03708.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. 2024. Weak-to-strong search: Align large language models via searching over small language models. arXiv preprint arXiv:2405.19262.

A Concept Accuracy

We provide the mean concept accuracies of Middle-EN, Elementary-EN and Elementary-ZH of the evaluated models across different concepts in Fig. 7, Fig. 8 and Fig. 9.

We illustrate the concept accuracies on Middle-ZH, Elementary-EN and Elementary-ZH for some selected models in Fig. 10, Fig. 11 and Fig. 12. For the results of all models, please refer to Fig. 13, Fig. 14, Fig. 15 and Fig. 16.

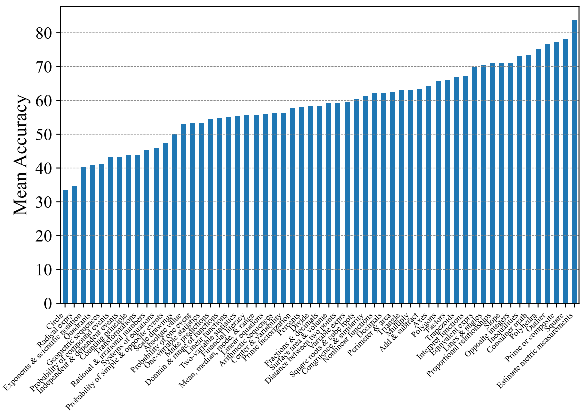


Figure 7: Mean concept accuracies of Middle-EN.

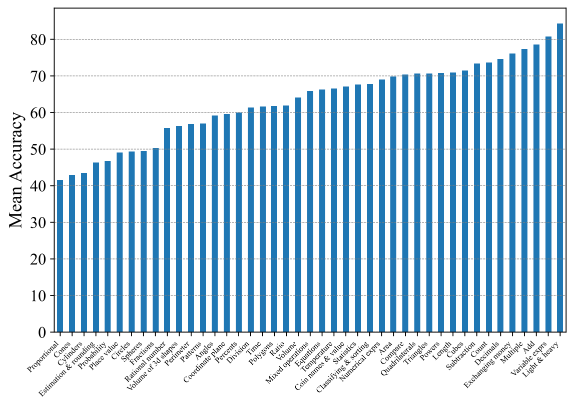


Figure 8: Mean concept accuracies of Elementary-EN.

B Details on the Evaluated Models

In this section, we offer a detailed overview of the Large Language Models (LLMs) and present the corresponding model links in Table 7. For enhanced safety and better conversational abilities, most LLMs undergo alignment techniques (Zhang et al., 2023; Shen et al., 2023; Zhou et al., 2023, 2024; Liu et al., 2024b) to align with human intents. We include instruction-finetuned models, math language models and chat models aligned with human intents in our evaluation.

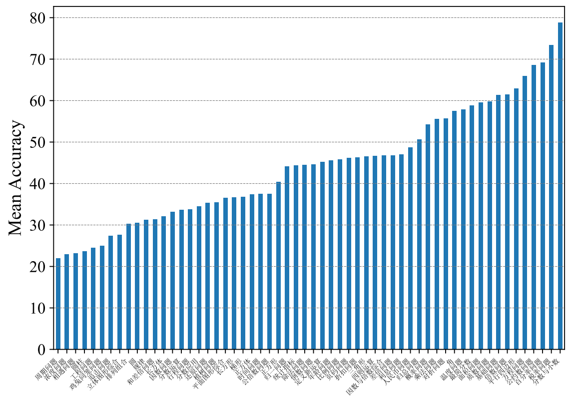


Figure 9: Mean concept accuracies of Elementary-ZH.

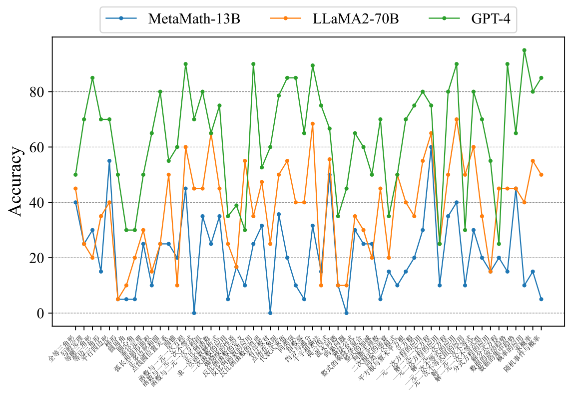


Figure 10: Concept accuracies on Middle-ZH.

- GPT-3.5/GPT-4 (OpenAI, 2023): The most powerful closed-model from OpenAI. We utilize its API: gpt-3.5-turbo and gpt-4.
- LLaMa2-7B/13B/70B (Touvron et al., 2023b): A set of open-source models developed by Meta.
- Qwen-14B/72B (Bai et al., 2023b): This model pre-trained on multilingual data, concentrates on Chinese and English languages. We employ both the Qwen-Base-14B, and the Qwen-Base-72B.
- Baichuan2-13B (Baichuan, 2023): This model demonstrates impressive performance in both Chinese and English benchmarks.
- MetaMath-13B (Megill and Wheeler, 2019): A domain-specific language model for mathematical reasoning, fine-tuned from the LLaMA-2 model using the MetaMathQA [5] dataset.

[5] `https://huggingface.co/datasets/meta-math/MetaMathQA`

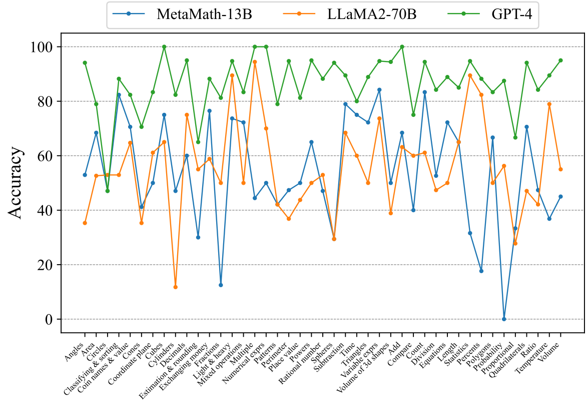


Figure 11: Concept accuracies on Elementary-EN.

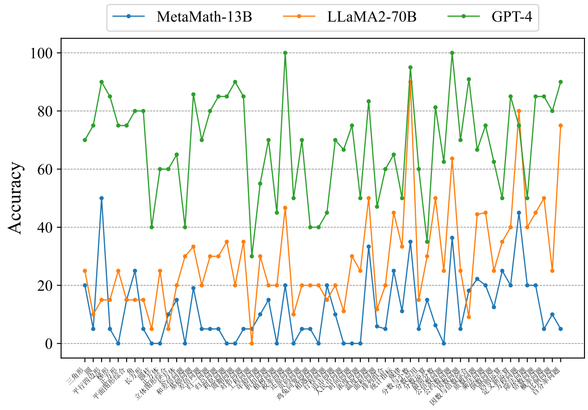


Figure 12: Concept accuracies on Elementary-ZH.

- WizardMath-13B (Luo et al., 2023): Another domain-specific language model for mathematical reasoning, fine-tuned from the LLaMA-2 model using reinforcement learning.

- MAmmoTH-13B (Yue et al., 2023): This model is specifically designed for general math problem-solving and has been fine-tuned from the LLaMA model using the MathInstruct [6] dataset. This dataset features training data that includes both chain-of-thought (CoT) and program-of-thought (PoT) rationales.

- Yi-6B/34B (Team, 2023b): This model released by 01 shows promising performance results in both Chinese and English.

- ChatGLM3-6B (Zeng et al., 2022): a lightweight and high-performance pre-trained dialogue model released by Zhipu AI in both Chinese and English.

- InternLM-7B/20B (Team, 2023a): A Multilingual Language Model with Progressively Enhanced Capabilities released by InternLM team.

- InternLM-Math-7B/20B (Team, 2023a): Well-performed math reasoning language models.

- DeepSeekMath-7B (Shao et al., 2024): One powerful mathematical language model released by DeepSeek.

C Classifying questions by concept

We compute the accuracy and Mean Reciprocal Rank[7] (MRR) of two closed-source LLMs (i.e., GPT-3.5 and GPT-4) and five open-source LLMs (LLaMA2-7B/13B, DeepSeekMath-7B, Qwen1.5-7B/14B) in classifying each question by concept.

The calculation method is different for closed-source and open-source methods. Specifically, for the open-source LLMs, we follow the benchmark processing on the MMLU dataset in the OpenCompass project (Contributors, 2023) to use options (i.e., A, B, ...) for labeling each concept and then calculate the log-likelihood for each option as the basis for ranking. For the closed-source LLMs (i.e., GPT-3.5 and GPT-4), as we cannot obtain the log-likelihood for each option and these closed-source methods have strong abilities to follow instructions, we use a system prompt to make the model output the relevant concepts the question belongs to, where the rankings are obtained from most to least relevant to the corresponding question. After getting the ranked concepts using the above methods, we calculate MRR and accuracy, which is shown in Table 9 and Table 8.

D Details on the ConceptMath

As shown in Table 10, Table 11, Table 12 and Table 13, we have provided the details on the three-level hierarchical system of our ConceptMath for better illustration.

E Details on 5-shot Prompts

We provide the 5-shot prompts for our ConceptMath in Pages 17-20.

[6]https://huggingface.co/datasets/TIGER-Lab/MathInstruct

[7]https://en.wikipedia.org/wiki/Mean_reciprocal_rank

Figure 13: Concept accuracies on Elementary-EN of more models.

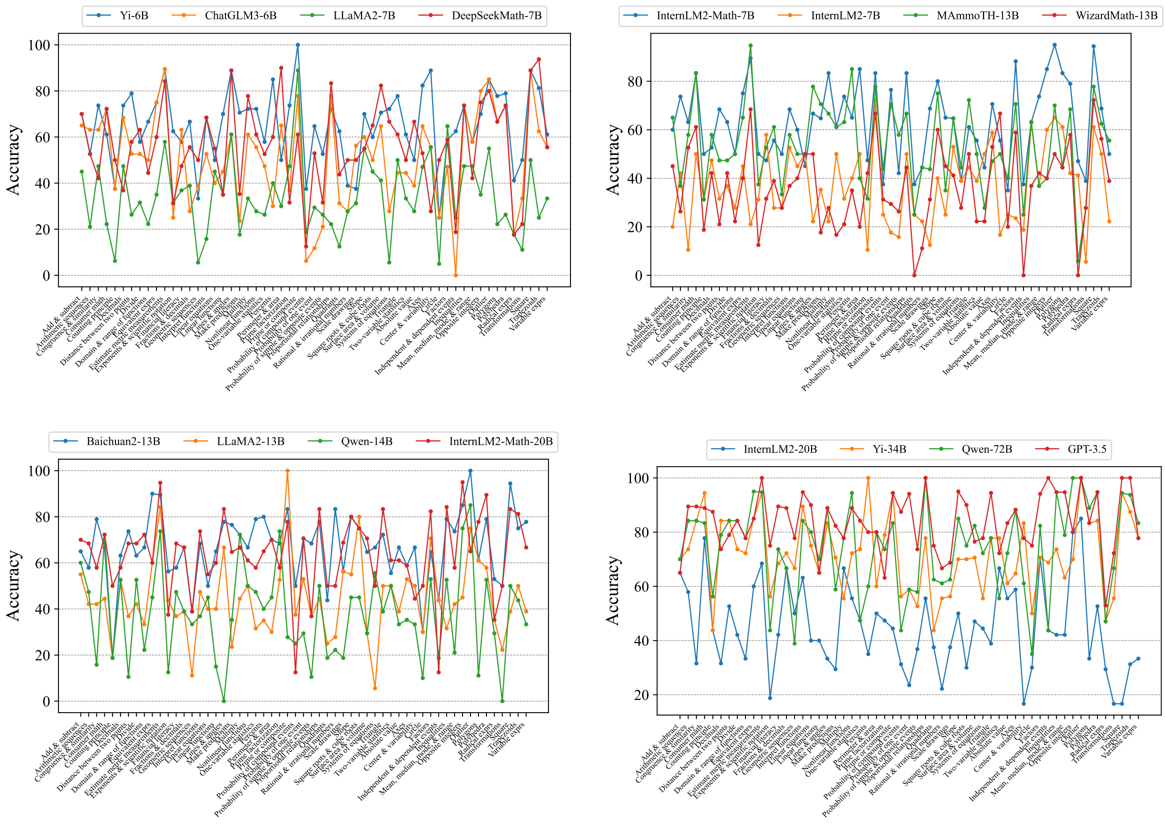


Figure 14: Concept accuracies on Middle-EN of more models.

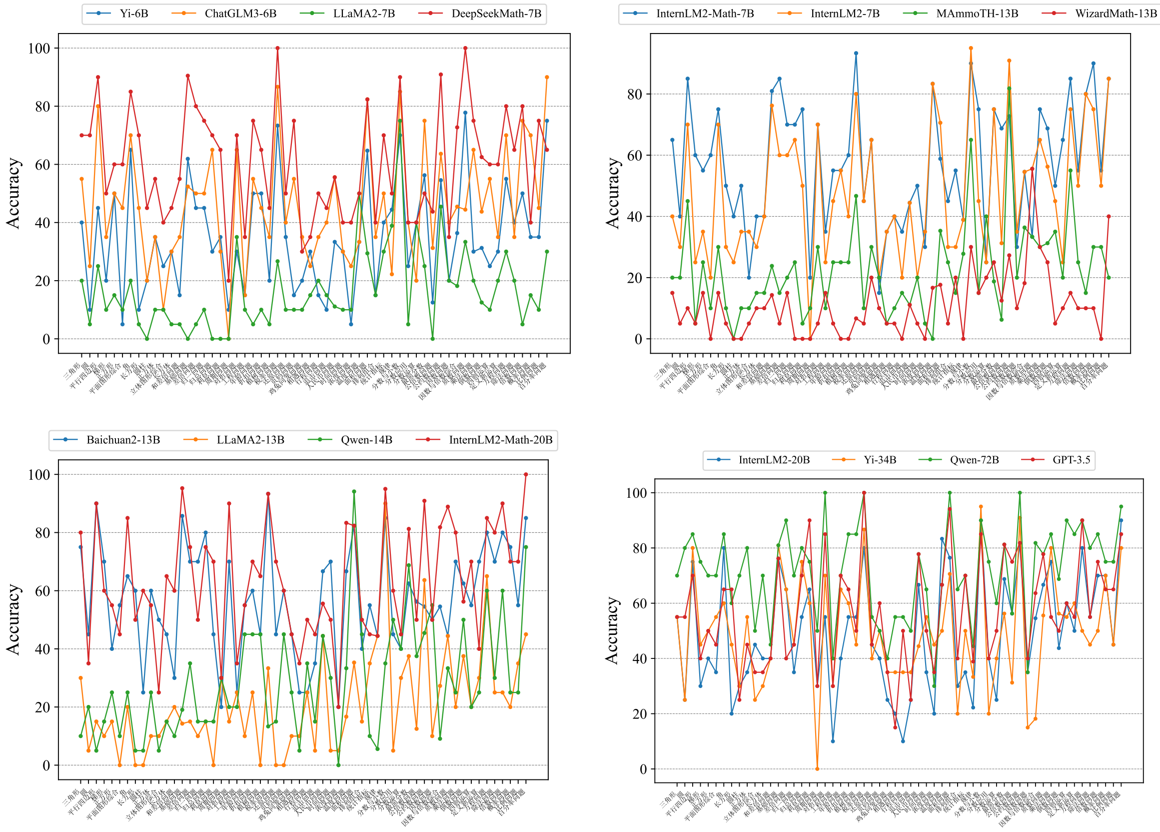


Figure 15: Concept accuracies on Elementary-ZH of more models.

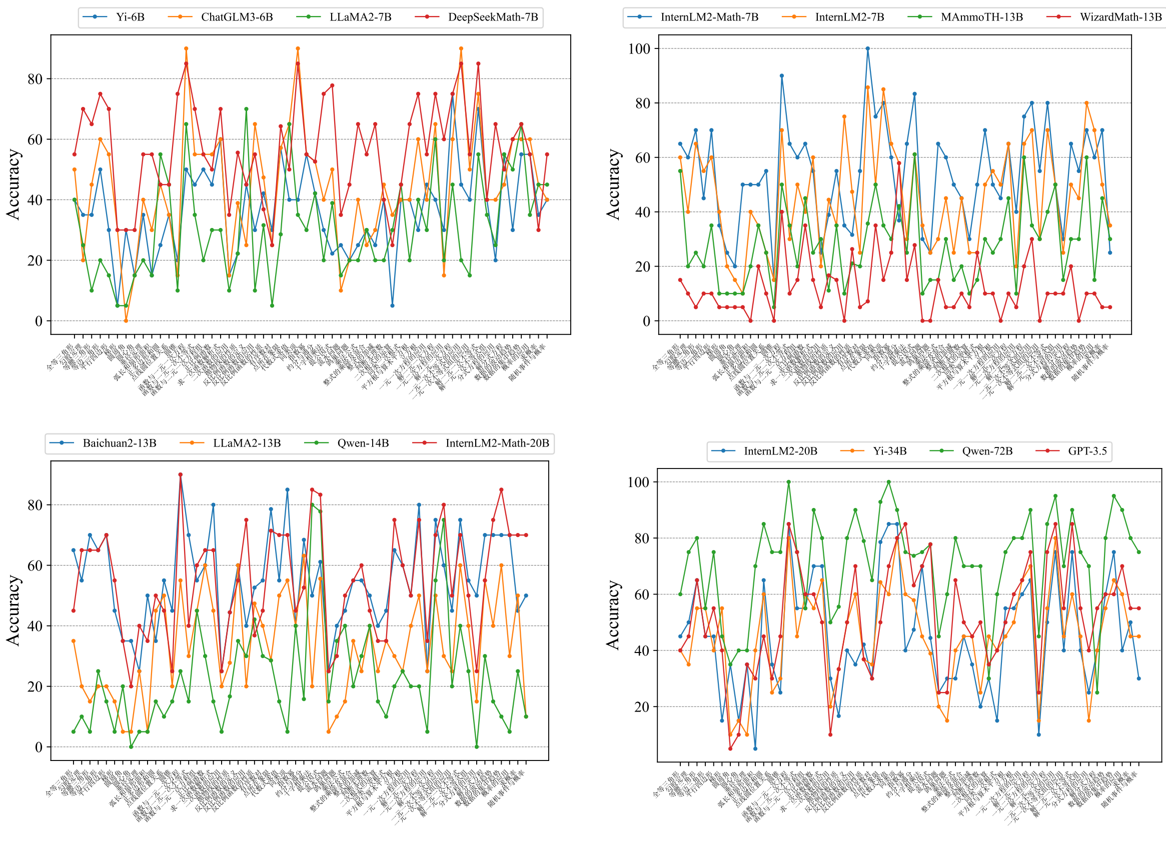


Figure 16: Concept accuracies on Middle-ZH of more models.

Models		HuggingFace Link / OpenAI Model
ChatGLM3	ChatGLM3-6B	https://huggingface.co/THUDM/chatglm3-6b
DeepSeekMath	DeepSeekMath-7B	https://huggingface.co/deepseek-ai/deepseek-math-7b-instruct
Baichuan2	Baichuan2-13B	https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat
MetaMath	MetaMath-13B	https://huggingface.co/meta-math/MetaMath-13B-V1.0
WizardMath	WizardMath-13B	https://huggingface.co/WizardLM/WizardMath-13B-V1.0
MAmmoTH	MAmmoTH-13B	https://huggingface.co/TIGER-Lab/MAmmoTH-13B
InternLM	InternLM-7B	https://huggingface.co/internlm/internlm2-chat-7b
	InternLM-20B	https://huggingface.co/internlm/internlm2-chat-20b
	InternLM-Math-7B	https://huggingface.co/internlm/internlm2-math-7b
	InternLM-Math-20B	https://huggingface.co/internlm/internlm2-math-20b
Yi	Yi-6B	https://huggingface.co/01-ai/Yi-6B-Chat
	Yi-34B	https://huggingface.co/01-ai/Yi-34B-Chat
LLaMA2	LLaMA2-7B	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
	LLaMA2-13B	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
	LLaMA2-70B	https://huggingface.co/meta-llama/Llama-2-70b-chat
Qwen	Qwen-14B	https://huggingface.co/Qwen/Qwen-14B-Chat
	Qwen-72B	https://huggingface.co/Qwen/Qwen-72B-Chat
GPT	GPT-3.5	gpt-3.5-turbo
	GPT-4	gpt-4

Table 7: Model links.

Model	Average	Elementart-EN	Middle-EN	Elementart-ZH	Middle-ZH
LLaMA2-7B	18.41	29.60	4.76	25.00	14.29
LLaMA2-13B	28.63	41.26	18.05	27.84	27.38
DeepSeekMath-7B	43.52	39.46	46.37	47.16	41.07
Qwen1.5-7B	42.32	47.98	48.12	38.07	35.12
Qwen1.5-14B	51.58	56.05	58.15	52.84	39.29
GPT-3.5	56.86	56.40	58.80	53.02	59.20
GPT-4	68.36	68.60	72.20	66.05	66.60

Table 8: The average accuracy of LLMs for classifying each question by concept.

Model	Average	Elementart-EN	Middle-EN	Elementart-ZH	Middle-ZH
LLaMA2-7B	32.91	43.81	17.56	41.35	28.90
LLaMA2-13B	44.35	56.99	32.97	46.60	40.82
DeepSeekMath-7B	57.73	55.48	59.98	61.09	54.36
Qwen1.5-7B	56.58	61.43	58.62	56.57	49.68
Qwen1.5-14B	64.69	70.20	69.06	64.83	54.65
GPT-3.5	69.11	73.95	68.82	64.56	69.10
GPT-4	78.17	81.45	80.14	75.80	75.27

Table 9: The average MRR of LLMs for classifying each question by concept.

Elementary-EN (5-shot prompts)

Q: For a party, Oscar gets 4 different kinds of soda. He buys a 6-pack of each kind. At home, he divides the sodas evenly among 3 coolers. How many sodas are in each cooler?
A: Oscar buys 4 different kinds of soda, and he gets a 6-pack of each kind. This means that he has 4 * 6 = 24 sodas in total. He divides these 24 sodas evenly among 3 coolers. So number of sodas per cooler is 24 / 3 = 8. Therefore, the answer is 8.
Q: Amazing Animals Zoo is building a new home for its 8 zebras. Each zebra will get 2 acres of grassy land to roam around in. The zebras will have a pond that covers 1 acre. How big will the zebras' new home be?
A: The total required grassland area for the zebras is: 8 zebras * 2 acres per zebra = 16 acres. Then, we need to add the extra 1 acre for the pond: 16 acres (for the zebras) + 1 acre (for the pond) = 17 acres. Therefore, the answer is 17.
Q: The state taxes personal income exceeding 3500 yuan (this portion not to exceed 3500 yuan) at a rate of 3%. If Xiaoming's father's monthly salary is 4700 yuan, how much personal income tax should Xiaoming's father pay each month?
A: To calculate the amount of personal income tax Xiaoming's father should pay, we can follow this simple process: Subtract the tax-free threshold from the monthly salary. Taxable income = Monthly salary - Tax-free threshold = 4700 - 3500 = 1200 yuan. Apply the tax rate to the amount that exceeds the threshold. Tax owed = Taxable income $*$ Tax rate = 1200 * 3% = 36 yuan. Therefore, the answer is 36.
Q: Find the 10th term in the arithmetic sequence where the first term is 3 and the common difference is 4.
A: To find the nth term in an arithmetic sequence, we can use the formula: $a_n = a_1 + (n-1) * d$, where: a_n is the nth term, a_1 is the first term, d is the common difference, n is the term number. Given that the first term (a_1) is 3 and the common difference (d) is 4, we can find the 10th term (a_{10}) by plugging these values into the formula: $a_{10} = a_1 + (10-1) * d = 3 + 9 * 4 = 3 + 36 = 39$. Therefore, the answer is 39.
Q: A regular polygon has an interior angle of 156 degrees. How many sides does the polygon have?
A: To find the number of sides of a regular polygon with a given interior angle, we can use the formula: Interior Angle = [(n - 2) * 180] / n, where n is the number of sides. Given that interior angle is 156: 156 = [(n - 2) * 180] / n. Solve for n, so we get n = 15. Therefore, the answer is 15.

Middle-EN (5-shot prompts)

Q: $g(x) = x^2$, $f(x) = 3(x-2)^2$, $h(x) = g(x) - f(x)$, $y = h(23)$, give me the value of y.
A: First, calculate $g(23)$ and $f(23)$: $g(23) = 23^2 = 529$. Next, calculate $f(23)$: $f(23) = 3 * (23 - 2)^2 = 3 * (21)^2 = 3 * 441 = 1323$. Now, we calculate $h(23)$: $h(23) = g(23) - f(23) = 529 - 1323 = -794$. Therefore, the answer is -794.
Q: What is the slope of the line segment between the two points: (3,1), (5,2)?
A: The slope of a line between two points (x_1, y_1) and (x_2, y_2) can be found using the formula: $slope(m) = (y_2 - y_1)/(x_2 - x_1)$. Given the two points (3, 1) and (5, 2), we can substitute these values into the formula: $slope(m) = (2-1)/(5-3) = 1/2 = 0.5$. Therefore, the answer is 1/2.
Q: In a party, there are 5 cupcakes remaining and 10 children. A mother inserts 10 pieces of paper into a hat, 5 with a cupcake image and 5 without. The children who draw the cupcake paper will receive a cupcake. If the first child draws a paper with a cupcake image, what is the probability that the second child will also draw a cupcake paper?
A: Initially, there are 5 cupcake papers and 5 non-cupcake papers in the hat, making a total of 10 papers. After the first child draws a cupcake paper, there are 4 cupcake papers and 5 non-cupcake papers left in the hat, with a total of 9 papers. The probability that the second child will draw a cupcake paper is then the number of cupcake papers left divided by the total number of papers left: Probability = Number of cupcake papers left / Total number of papers left = 4/9. Therefore, the answer is 4/9.
Q: What is the total area of an irregular polygon that consists of two squares and two rectangles where the sides of the squares are 12 cm and 16 cm respectively, and the dimensions of the rectangles are 13 cm by 20 cm and 10 cm by 7 cm respectively?
A: To find the total area of an irregular polygon consisting of two squares and two rectangles, you would sum up the individual areas of each shape. The area of a square is given by the formula $A = s^2$, where s is the length of a side of the square. For the first square with a side of 12 cm: $A_1 = 12^2 = 144\, cm^2$. For the second square with a side of 16 cm: $A_2 = 16^2 = 256\, cm^2$. The area of a rectangle is given by the formula ($A = l * w$), where l is the length and w is the width of the rectangle. For the first rectangle with dimensions 13 cm by 20 cm: $A_3 = 13 * 20 = 260\, cm^2$. For the second rectangle with dimensions 10 cm by 7 cm: $A_4 = 10 * 7 = 70\, cm^2$. $Total\ Area = A_1 + A_2 + A_3 + A_4 = 144 + 256 + 260 + 70 = 730\, cm^2$. Therefore, the answer is 730.
Q: Susie is making invitations to her party this coming weekend. It takes her about 5 minutes to make each invitation. It took her 2 hours to finish all the invitations. How many people is Susie inviting?
A: First, convert the 2 hours into minutes to match the unit of time used for the invitation-making process. 2 hours = 120 minutes. Then, divide the total time spent by the time it takes to make one invitation to find out how many invitations she made: 120 minutes total / 5 minutes per invitation = 24 invitations. Therefore, the answer is 24.

Elementary-ZH (5-shot prompts)

Q: 小文用量角器测量一个角的度数时，误把外圈刻度看成了内圈刻度，量出的角是 65°，这个角实际是多少度?
A: 量角器一般有两个圈的刻度，内圈刻度通常是顺时针方向从 0 度到 180 度，外圈刻度通常是逆时针方向从 0 度到 180 度。如果小文误把外圈刻度看成了内圈刻度，那么测量出的角度其实是从 180° 减去错误读数。因此，实际的角度应该是：$180° - 65° = 115°$。因此，答案是 115。
Q: 小明以每小时 2 千米的速度上山，到了山顶后立即以每小时 3 千米的速度按原路返回。求小明上下山的平均速度是多少千米?
A: 要计算小明上下山的平均速度，我们需要知道他上山和下山所走的总距离以及总时间。首先，总距离 = 上山距离 + 下山距离 = d + d = 2d 千米。然后，计算总时间：上山时间 = 上山距离/上山速度 = d / 2 小时，下山时间 = 下山距离/下山速度 = d / 3 小时，因此，总时间 = 上山时间 + 下山时间 = d / 2 + d / 3 小时。最后，计算平均速度：平均速度 = 总距离 / 总时间 = 2d / (d/2 + d/3) = 2d / (5d / 6) = 12 / 5 = 2.4 千米/小时。因此，答案是 2.4 千米/小时。
Q: 一个礼盒长 20cm、宽 15cm、高 10cm，用彩带捆扎这样的一个礼盒，至少需要多长的彩带?（接头处 12 厘米）
A: 要捆扎这样一个礼盒，通常情况下我们会沿着礼盒的长和宽方向各绕一圈，即绕过礼盒的长和宽两次，然后再留一些长度作为接头。首先，计算沿着礼盒的长方向绕一圈的长度：绕长方向长度 = (礼盒的长 + 礼盒的高) * 2 = (20 + 10)* 2 = 30 * 2 = 60 cm 然后，计算沿着礼盒的宽方向绕一圈的长度：绕宽方向长度 = (礼盒的宽 + 礼盒的高) * 2 = (15 + 10)* 2 = 25 * 2 = 50 cm 最后，计算总需要彩带长度：总长度 = 绕长方向长度 + 绕宽方向长度 + 接头长度 = 60 + 50 + 12 = 122 cm 。 因此，答案是 122 厘米。
Q: 千岛湖大桥的一侧每隔 10 米安装 1 盏灯，共需 91 盏（两端都装），现在改装 61 盏（两端都装），平均每相邻两盏灯之间的距离是多少米?
A: 原来每隔 10 米安装 1 盏灯，共需 91 盏，这意味着桥的总长度覆盖了 90 个 10 米的间隔，所以桥的总长度 = 90 间隔 * 10 米/间隔 = 900 米。现在改装 61 盏灯（两端都装），也就是说将这 900 米分成 60 个间隔（因为 61 个灯包含了两端的灯，所以间隔数比灯的数量少一个），每个间隔的长度 = 900 米 / 60 间隔 = 15 米/间隔。因此，答案是 115 米。
Q: a 是 b 的 3/4，则 a 比 b 少 ( )，b 比 a 多 ( )，b 与 a 的比是 ( )，如果 a、b 两数的和是 210，则 a ＝ ( )，b ＝ ( )，括号处分别为多少?
A: 已知 a 是 b 的 3/4，则 a = 3/4 * b。因此，a 比 b 少 (1-3/4)/1 = 1/4，b 比 a 多（1-3/4)/(3/4) = 1/3，b 与 a 的比为 b : (3/4 * b) = 4:3。 若已知 a、b 两数的和是 210，则 a+b = 210，将 a = (3/4) * b 代入方程得到：3/4 * b + b = 210，解得 b = 120， a = 3/4 * b = 3/4 * 120 = 90。因此，答案分别是 1/4，1/3，4:3，90，120。

Middle-ZH (5-shot prompts)

Q: 求不等式 $-x^2-2x+8>0$ 的解集?
A: $-x^2-2x+8>0$, 通过分解因式方式，得到 $-(x-2)(x+4)>0$ 因此，答案是 (-4, 2)。
Q: 已知函数 $y=2x+3$ 与的图象 $y=-1/2x+m$ 交 y 轴于点 A、B，两函数图象的交点为 P(n,-2)，求 m 和 n 的值?
A: 两函数图像交点为 P(n,-2)，意味着这个点同时满足上述两个函数方程：将 P 点的坐标 (n,-2) 代入函数 $y=3x+3$，即 $-2=3n+3$，解得 $n=-5/2$。将 P 点坐标 (-5/2,-2) 代入函数 $y=-1/2x+m$，即 $-2=-1/2*(-5/2)+m$，解得 m=-13/4。因此，答案是 m 为-13/4，n 为-5/2。
Q: 已知 PA 是圆 O 的切线，A 为切点，PC 与圆 O 相交于 B、C 两点，PB=2cm，BC=8cm，则 PA 的长是多少厘米?
A: 由于 PA 是圆 O 的切线，根据切割线定理，我们知道从圆外一点引圆的两条割线，这一点到每条割线与圆的交点的两条线段长的积相等，即 $PA^2=PC*PB$, 已知 PB=2cm、BC=8cm，则 $PC=PB+BC=2cm+8cm=10cm$。$PA^2=PC*PB=10*2=20$，解得 PA = $2\sqrt{5}$cm。因此，答案是 $2\sqrt{5}$ 厘米。
Q: 南宋著名数学家杨辉精研数学，著有《详解九章算法》，对数的运算进行了深入研究与总结，类比其中的思想方法，可以解决很多数与式的计算问题已知 a，b 为实数，且 $a+b=3, ab=1$, 计算可得: $a^2+b^2=7$, $a^3+b^3=18$, $a^4+b^4=47$, 由此求 a^5+b^5 得多少?
A: 已知 $a+b=3$ 和 $ab=1$，可得: $a^2+b^2=(a+b)^2-2ab=3^2-2*1=9-2=7a^3+b^3=(a+b)(a^2-ab+b^2)=3(7-1)=3*6=18a^4+b^4=(a^2+b^2)^2-2a^2b^2=7^2-2*1^2=49-2=47a^5+b^5=(a+b)(a^4-a^3b+a^2b^2-ab^3+b^4)=3(a^4+b^4-a^2b^2)=3(47-1)=3*46=138$。因此，答案是 138。
Q: 某单位要招聘 1 名英语翻译，甲、乙两人报名参加了听、说、读、写等项素质测试，甲的成绩分别为：90 分、80 分、85 分、78 分；乙的成绩分别为：78 分、82 分、85 分、88 分．如果把听、说、读、写的成绩按 2 : 2 : 3 : 3 计算素质测试平均成绩，那么甲和乙的平均成绩分别是多少?
A: 甲的平均成绩为 $(90*2+80*2+85*3+78*3)/(2+2+3+3)=(180+160+255+234)/10=829/10=82.9$ 分，乙的平均成绩为 $(78*2+82*2+85*3+88*3)/(2+2+3+3)=(156+164+255+264)/10=839/10=83.9$ 分。因此，答案是甲的平均成绩是 82.9 分，乙的平均成绩是 83.9 分。

LEVEL1	LEVEL2	LEVEL3	# Samples
Calculation & Properties	Calculation	Add	19
		Decimals	20
		Division	19
		Equations	18
		Fractions	16
		Mixed Operations	18
		Multiple	18
		Numerical Expressions	20
		Place Value	16
		Powers	20
		Rational Number	17
		Subtraction	19
		Variable Expressions	19
	Properties	Compare	20
		Count	18
		Estimation & Rounding	20
		Patterns	19
Geometry	Angles	Angles	17
	Coordinate Plane	Coordinate Plane	18
	Three-dimensional Shapes	Cones	17
		Cubes	20
		Cylinders	17
		Spheres	17
		Volume of 3D shapes	18
	Two-dimensional Shapes	Circles	17
		Perimeter	19
		Polygons	18
		Quadrilaterals	17
		Triangles	18
Measurement	Basic Knowledge	Temperature	19
		Time	20
	Money	Coin Names & Value	17
		Exchanging Money	17
	Ratio	Percent	17
		Proportion	18
		Ratio	19
	Size	Area	19
		Length	20
		Volume	20
	Weight	Light & Heavy	20
Statistics	Classifying & Sorting	Classifying & Sorting	17
	Data	Mode/Mean/Median/Range	19
	Probability	Probability	16

Table 10: Details of the hierarchical concepts in Elementary-EN.

LEVEL1	LEVEL2	LEVEL3	# Samples
Calculation	Basic Calculation	Add & Subtract	20
		Decimals	19
		Divide	19
		Exponents & Scientific Notation	16
		Fractions & Decimals	18
		Multiply	18
		Square Roots & Cube Roots	20
	Consumer Math	Consumer Math	18
	Financial Literacy	Financial Literacy	19
	Integers	Absolute Value	18
		Opposite Integers	20
	Measurement	Measurement Metric	19
	Number Theory	Factors	20
		Prime Factorization	19
		Prime or Composite	18
	Percents	Percents	20
	Rational & Irrational Numbers	Rational & Irrational Numbers	18
	Ratios & Rates	Proportional Relationships	18
	Sequences	Arithmetic Sequences	19
		Geometric Sequences	18
Expressions, equations, and functions	Equations	Linear Equations	20
		Systems of Equations	18
	Expressions	Equivalent Expressions	20
		Radical	17
		Variable	18
	Function	Domain & Range of Functions	18
		Interpret Functions	19
		Linear Functions	20
		Nonlinear Functions	18
	Inequalities	Inequalities	19
Geometry	Congruence & Similarity	Congruence & Similarity	19
	Coordinate Plane	Axes	17
		Distance Between Two Points	19
		Quadrants	16
	Scale Drawings	Scale Drawings	16
	Slope	Slope	20
	Three-dimensional Figures	Polyhedra	19
		Surface Area & Volume	17
	Transformations	Transformations	18
	Two-dimensional Figures	Circle	20
		Lines & Angles	18
		Perimeter & Area	20
		Polygons	18
		Square	18
		Trapezoids	16
		Triangle	18
Statistic and Probability	Data	Center & Variability	18
		Mean, Median, Mode & Range	19
		Outlier	20
	One-variable Statistics	One-variable Statistics	19
	Probability	Counting Principle	16
		Independent & Dependent Events	16
		Make Predictions	17
		Probability of Compound Events	16
		Probability of One Event	17
		Probability of Simple and Opposite Events	19
	Two-variable Statistics	Two-variable Statistics	18

Table 11: Details of the hierarchical concepts in Middle-EN.

LEVEL1	LEVEL2	LEVEL3	# Samples
几何 (Geometry)	平面图形 (Two-dimensional shapes)	三角形 (Triangles)	20
		圆 (Circle)	20
		平行四边形 (Parallelogram)	20
		梯形 (Trapezium)	20
		正方形 (Square)	20
		平面图形综合 (Synthesis Problem)	20
		角 (Angle)	20
		长方形 (Rectangle)	20
	立体图形 (Three-dimensional Shapes)	圆柱 (Cylinder)	20
		正方体 (Cube)	20
		立体图形综合问题 (Synthesis Problem)	20
		长方体 (Cuboid)	20
应用 (Application)	基础 (Fundamental Problem)	和差倍问题 (Add & Differential & Multiple)	20
		基础 (Basics)	21
		差倍问题 (Differential)	20
		归一问题 (Normalization)	20
		归总问题 (Induction)	20
	经典问题 (Classical Problem)	利息问题 (Interest)	20
		周期问题 (Period)	10
		对折问题 (Folding)	20
		工程问题 (Engineering)	20
		年龄问题 (Age)	20
		折扣问题 (Discount)	20
		植树问题 (Planting)	20
		税率问题 (Tax)	15
		还原问题 (Reduction)	20
		页码问题 (Pagination)	20
		鸡兔同笼问题 (Chickens & Rabbits in the Same Cage)	20
	路程问题 (Distance Problem)	相遇问题 (Encounter)	20
		行程问题 (Travel)	20
		追击问题 (Pursuit)	20
度量与统计 (Measurement and Statistics)	度量 (Measurement)	人民币问题 (RMB)	9
		时间问题 (Time)	20
		浓度问题 (Concentration)	20
		温度问题 (Temperature)	6
		面积问题 (Area)	17
	统计 (Statistics)	排列组合 (Permutation)	20
		统计指标 (Statistical Metrics)	20
		规律 (Law)	18
数与代数 (Number and algebra)	分数运算 (Fractional Operation)	分数与小数 (Fraction & Decimal)	20
		分数应用 (Fractional Application)	20
		分数运算 (Fractional Operation)	20
		最简分数 (Simplest Fraction)	16
	因数与倍数 Factors & Multiples	公倍数问题 (Common Multiples)	16
		公约数问题 (Common Divisors)	11
		因数问题 (Factor)	20
		因数与倍数综合问题 (Synthesis Problem)	11
		质数问题 (Prime Number)	9
	基础运算 (Basic Operation)	乘法问题 (Multiplication)	20
		倒数问题 (Reciprocal Problem)	16
		四则运算 (Four-rule Operation)	20
		新运算定义 (New Operation Definition)	20
		方程问题 (Equation)	20
		除法问题 (Division)	20
	比 (Ratio)	倍数问题 (Multiple)	20
		概率问题 (Probability)	20
		比例问题 (Proportion)	20
		百分率问题 (Percentage)	20

Table 12: Details of the hierarchical concepts in Elementary-ZH.

LEVEL1	LEVEL2	LEVEL3	# Samples
儿何 (Geometry)	三角形 (Triangle)	全等三角形 (Congruent Triangle)	20
		勾股定理 (Pythagorean Theorem)	20
		等腰三角形 (Isosceles Triangle)	20
		等边三角形 (Equilateral Triangle)	20
	四边形 (Quadrilateral)	平行四边形 (Parallelogram)	20
		梯形 (Trapezium)	20
	圆 (Circle)	圆周角 (Angle of Circumference)	20
		圆心角 (Angle of Center)	20
		垂径定理 (Vertical Path Theorem)	20
		弧长和扇形面积 (Arc length & Sector Area)	20
		正多边形和圆 (Regular Polygons & Circles)	20
		点线圆位置关系 (Relations of Point, Line & Circle)	20
	立体图形 (Three-dimensional Shapes)	圆锥 (Cone)	20
函数 (Function)	一次函数 (Linear Function)	函数与一元一次方程 (Univariate Function & Equation)	20
		函数与一元一次不等式 (Linear Functions & Univariate Linear Inequalities)	20
		一次函数与二元一次方程组 (Linear Functions & System of Binary Linear Equations)	20
		正比例函数 (Proportional Function)	20
		一次函数解析式 (Analytical Formula of Linear Functions)	20
	二次函数 (Quadratic Function)	二次函数的应用 (Applications of Quadratic Functions)	20
		抛物线的性质 (Properties of Parabolas)	18
	反比例函数 (Inverse Proportional Function)	定义 (Definition)	20
		应用 (Applications)	20
		性质 (Properties)	19
	平面直角坐标系 (Rectangular Coordinate System)	有序数对 (Ordered Pair)	20
		象限中的点 (Points of Quadrant)	14
数与式 (Number and Expression)	代数式 (Algebra Expression)	代数式求值 (Algebraic Expression Evaluation)	20
		同类项 (Similar Items)	20
	分式 (Fraction)	指数幂 (Exponential Power)	20
		约分 (Fraction Reduction)	19
	因式 (Factor)	十字相乘法 (Cross Multiplication)	20
		公因式提取 (Common Factor Extraction)	18
	应用 (Application)	流水问题 (Flow Problem)	20
		鸽巢问题 (Pigeon Nest Problem)	20
	整式 (Integral Expression)	乘法公式 (Multiplication)	20
		整式的乘除及混合 (Multiplication, Division & Mixing)	20
		整式的加减 (Addition & Subtraction)	20
	无理数 (Irrational Number)	无理数识别 (Irrational Number Recognition)	20
	根式 (Radical Expression)	二次根式的运算 (Operation of Quadratic Radicals)	20
		同类二次根式 (Similar Quadratic Radicals)	20
		平方根与算术平方根 (Square Root & Arithmetic Square Root)	20
		立方根 (Cube Root)	20
方程与不等式 (Equations & Inequalities)	一元一次方程 (Linear Equation in One Variable)	一元一次方程的应用 (Applications)	20
		解一元一次方程 (Solutions)	20
	一元二次方程 (Quadratic Equation in One Variable)	一元二次方程的应用 (Applications)	20
		解一元二次方程 (Solutions)	20
	不等式与不等式组 (Inequalities & Groups of Inequalities)	一元一次不等式的应用 (Applications of Unary First Order Inequality)	20
		一元一次不等式组的应用 (Applications of Unary First Order Groups of Inequalities)	20
		解一元一次不等式 (Solve the First Inequality of One Variable)	20
		解一元一次不等式组 (Solve Unary First Order Groups of Inequalities)	20
	分式方程 (Fractional Equation)	分式方程的应用 (Application of Fractional Equation)	20
		解分式方程 (Solve Fractional Equation)	20
统计与概率 (Statistics and Probability)	数据分析 (Data Analysis)	数据的波动趋势 (Fluctuating Trend of Data)	20
		数据的集中趋势 (Central Tendency of Data)	20
	概率 (Probability)	概率的应用 (Applications of Probability)	20
		求概率 (Find Probability)	20
		随机事件与概率 (Random Events & Probabilities)	20

Table 13: Details of the hierarchical concepts in Middle-ZH.