Front. Digit. Educ., 2025, 2(2): 16
https://doi.org/10.1007/s44366-025-0053-z

RESEARCH ARTICLE

MathEval: A Comprehensive Benchmark for Evaluating Large Language Models on Mathematical Reasoning Capabilities

Tianqiao Liu[a,b], **Zui Chen**[b], **Zhensheng Fang**[b], **Weiqi Luo**[a], **Mi Tian**[b], **Zitao Liu**[a]

[a] Guangdong Institute of Smart Education, Jinan University, Guangzhou 510632, China

[b] TAL Education Group, Beijing 102206, China

Abstract Mathematical reasoning is a fundamental aspect of intelligence, encompassing a spectrum from basic arithmetic to intricate problem-solving. Recent investigations into the mathematical abilities of large language models (LLMs) have yielded inconsistent and incomplete assessments. In response, we introduce MathEval, a comprehensive benchmark designed to methodically evaluate the mathematical problem-solving proficiency of LLMs in various contexts, adaptation strategies, and evaluation metrics. MathEval consolidates 22 distinct datasets, encompassing a broad spectrum of mathematical disciplines, languages (including English and Chinese), and problem categories (ranging from arithmetic and competitive mathematics to higher mathematics), with varying degrees of difficulty from elementary to advanced. To address the complexity of mathematical reasoning outputs and adapt to diverse models and prompts, we employ GPT-4 as an automated pipeline for answer extraction and comparison. Additionally, we trained a publicly available DeepSeek-LLM-7B-Base model using GPT-4 results, enabling precise answer validation without requiring GPT-4 access. To mitigate potential test data contamination and truly gauge progress, MathEval incorporates an annually refreshed set of problems from the latest Chinese National College Entrance Examination (Gaokao-2023, Gaokao-2024), thereby benchmarking genuine advancements in mathematical problem solving skills.

Keywords mathematical reasoning, large language models, benchmark, answer grading

1 Introduction

Mathematics has long been regarded as a cornerstone of human intelligence, encompassing a wide spectrum of competencies ranging from elementary arithmetic to complex logical inference (Ahn et al., 2024; Liu et al., 2024). In recent years, the rapid proliferation of research on large language models (LLMs) has significantly augmented these models' aptitude for mathematical reasoning (Chen et al., 2025; Ma et al., 2024). Such advancements show considerable promise not only for theoretical progress but also for practical applications, particularly in the context of education (Chen et al., 2024; Zhan et al., 2024; Zheng et al., 2024). By providing on-demand, individualized explanations and step-by-step guidance, LLMs with sophisticated mathematical and reasoning capabilities can effectively scaffold student learning. They illustrate structured approaches to problem solving, breaking down intricate questions into manageable components and thereby fostering deeper comprehension. Beyond practice exercises and targeted feedback, these models serve as catalysts for more rigorous cognitive engagement, nudging students toward refined reasoning pathways. The stronger a model's math reasoning abilities, the more effective it becomes in educational settings. With numerous LLMs available, selecting one with robust mathematical capabilities is crucial. However, the evaluation of these models remains challenging due to three primary issues: incomprehensiveness, inadequate adaptation to varying model types and datasets, and inconsistency.

Incomprehensiveness indicates that evaluations often do not cover a wide array of datasets, neglecting factors such as language diversity, problem types, and complexity levels. This limited scope can skew perceptions of a model's versatility and effectiveness. Inadequate adaptation highlights the shortcomings in current evaluations to flexibly accommodate different types of models and datasets. For instance, chat models, which have been fine-tuned during the alignment phase, are especially sensitive to the structure of

Received January 5, 2025; revised February 18, 2025; accepted March 3, 2025

Zitao Liu (✉)
E-mail: liuzitao@jnu.edu.cn

prompts. Similarly, evaluations should also adapt prompts to fit the specific characteristics of each dataset. For example, multiple-choice problems may require prompts that include hints to guide the selection from provided options, whereas math word problems (MWPs) might benefit from prompts that encourage chain-of-thought (CoT) reasoning. Inconsistency arises when the same model yields different performances on identical datasets, complicating the accurate estimation of its true capabilities. This issue primarily stems from the difficulty in verifying answers to MWPs, where outputs may include reasoning steps, equations, and final answers in various formats (e.g., 1/2 and (\frac{1}{2})). Extending this to different models and various types of datasets further complicates the evaluation. Rule-based methods for extracting and comparing answers, commonly utilized in benchmarks such as OpenCompass (OpenCompass, 2025) and HELM (Liang et al., 2023), often lack robustness. Even minor modifications can significantly alter the evaluation outcomes, making it impractical to tailor these rules for each specific model and dataset. Consequently, standardizing the process of extracting and comparing outputs continues to pose a significant challenge in benchmark evaluations. More related works are discussed in Section 2.

To address these challenges, we introduce MathEval, a comprehensive and unified benchmarking framework, as illustrated in Figure 1. MathEval incorporates 22 datasets in both Chinese and English, covering a wide range of mathematical problems from primary to high school levels, and includes a dynamically updated dataset to prevent test data contamination. Each dataset is meticulously categorized; for instance, the classic GSM8K (Cobbe et al., 2021) dataset represents the math scenario of English, MWPs, and primary school tasks. To tackle the adaptation challenge, MathEval employs tailored prompts suitable for various models and datasets, ranging from zero-shot to few-shot settings. This ensures a thorough assessment of each model's capabilities across diverse problem sets, promoting a fair comparison of mathematical abilities across models. MathEval leverages GPT-4 for both answer extraction and comparison, thereby avoiding the complexities of regular expression rules and setting a consistent evaluation standard. We have validated GPT-4's effectiveness by comparing its outputs against human-annotated result, with only minimal discrepancies noted. To our knowledge, this is the first comprehensive benchmark specifically designed to evaluate the mathematical capabilities of LLMs holistically. We have evaluated 52 models across 22 datasets under varied adaptation conditions, making the results publicly accessible.

Contributions of MathEval are outlined as follows:

- MathEval provides an extensive benchmark that includes a diverse array of mathematical problems across different types and difficulty levels. This thorough categorization facilitates detailed analyses that can unveil new insights and directions for future research in the field of LLMs and mathematical reasoning.
- We have developed a standardized method for comparing answers that effectively addresses the complexities associated with outputs from MWPs. For broader accessibility, we also offer a self-developed compare-answer model for researchers and developers who do not have access to GPT-4.
- Recognizing the potential for data contamination in LLM evaluations, MathEval implements a strategy of using a dynamically updated dataset. This approach ensures that the evaluation reflects the true, unlearned capabilities of LLMs in solving mathematical problems, providing a more accurate measure of their realistic mathematical reasoning ability.

2 Related Work

General benchmarks provide a comprehensive evaluation of LLMs and are widely used across various natural language understanding tasks to assess their

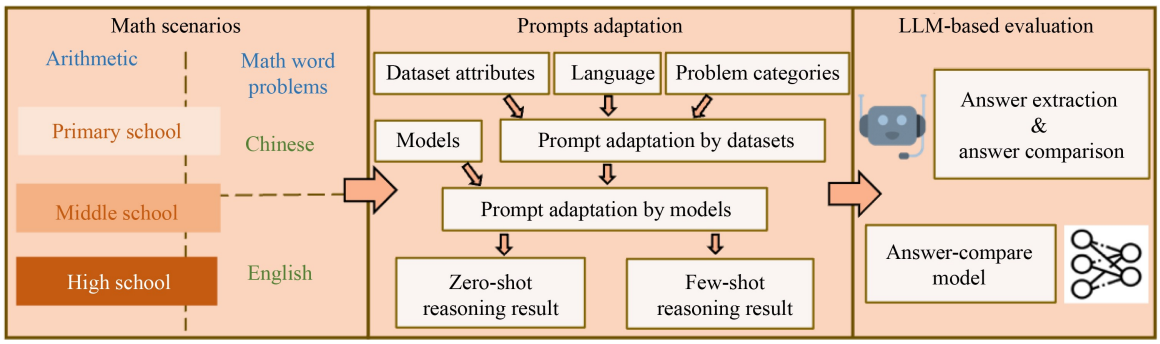

Figure 1 Three core components of MathEval addressing key challenges. LLM: large language model.

performance. The MMLU (Hendrycks et al., 2021a) benchmark is notable for its extensive collection of 57 tasks covering diverse domains, offering comprehensive challenges across varying subjects and levels of complexity. However, it was primarily focused on multiple-choice questions, which might not fully capture the depth of a model's reasoning capabilities compared to tasks requiring detailed solutions. AGIEval (Zhong et al., 2024) is centered on standardized exams like the SAT, LSAT, and GRE, testing models' reasoning, problem-solving, and language comprehension skills. The broader BIG-bench (Srivastava et al., 2023) initiative includes a diverse set of tasks designed to probe models on novel and complex linguistic capabilities, challenging them to demonstrate their robustness and versatility in a wide array of cognitive tasks beyond traditional benchmarks.

Domain-specific benchmarks are crucial for evaluating how well LLMs handle specialized tasks requiring deep field knowledge. HaluEval (Li et al., 2023b) assesses hallucination detection in LLMs using annotated samples, revealing that models frequently generate unverifiable information. LongBench (Bai et al., 2023b) tests long-context comprehension in English and Chinese across 21 datasets, showing that expanding context windows and enhancing memory mechanisms improve long-sequence understanding.

To the best of our knowledge, there is currently no comprehensive mathematical evaluation benchmark. A similar mathematical benchmark, Lila (Mishra et al., 2022), focuses on extending datasets by collecting task instructions and solutions as Python programs and then exploring some models' out-of-domain capabilities. A comprehensive benchmark for assessing the mathematical abilities of various models remains absent.

3 MathEval

In this section, we will delve into the essential aspects of MathEval's implementation by elaborating on its three main components: math scenarios, prompt adaptations, and evaluation methods. Finally, we will introduce the entire pipeline to provide a comprehensive understanding of how these components integrate to form MathEval.

3.1 | Math Scenarios

Figure 2 presents MathEval's compilation of 22 math datasets utilized in leading conference papers since 2010, spanning six scenarios across problem types (MWPs and arithmetic), languages (Chinese and English), and educational levels (primary school, middle school, and high school). This organization ensures comprehensive coverage of various math scenarios for robust evaluation. Notably, MathEval uniquely features the Arith3K, Gaokao-2023, Gaokao-2024, TAL-SCQ5K-EN, and TAL-SCQ5K-CN datasets, which are new additions not previously included in other benchmarks. Specifically, within the problem type dimension of our MathEval benchmark, three datasets, Arith3K, Big-Bench-Hard (BBH)-Math, and Math401, focus solely on arithmetic problems, while the remaining 19 datasets are dedicated to MWPs. For the language dimension, it is important to note that only MWPs' datasets require language categorization, and these are nearly evenly split with 10 in English and 9 in Chinese. Regarding the educational level, 12 datasets target primary school, 9 cater to high school, and only Arith3K is designed for middle school students. Predominantly, the datasets focus on primary school level English MWP, followed by primary and high school level Chinese MWP scenarios. Detailed information about each dataset is available in Subsection 4.2.

Building on this overview, we introduce some new datasets that have not been used by other benchmarks in MathEval: Arith3K, TAL-SCQ5K-EN, TAL-SCQ5K-CN, Gaokao-2023, and Gaokao-2024, each offering unique characteristics and challenges to the benchmark. These additions provide unique

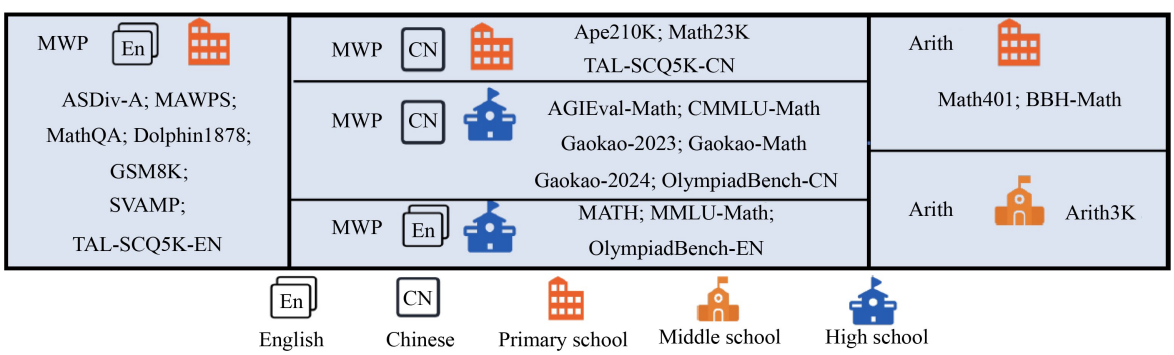

Figure 2 Overview of 22 datasets used in MathEval Framework. Arith: arithmetic; MWP: math word problem; BBH: Big-Bench-Hard.

characteristics and challenges that enhance the evaluation landscape in significant ways. The Gaokao datasets (Gaokao-2023 and Gaokao-2024) offer problems derived from actual national college entrance exams, providing a higher level of difficulty and realism. Additionally, the inclusion of TAL-SCQ5K-EN and TAL-SCQ5K-CN introduces large-scale multilingual MWP datasets, enabling more comprehensive evaluations across different languages. These enhancements collectively expand the scope and depth of mathematical problem types, educational levels, and linguistic diversity covered by MathEval, making it a more robust and versatile benchmark compared to MATH, GSM8K, and AGIEval-Math. For detailed descriptions of these datasets, please refer to Subsection 4.1.

3.2 | Prompts Adaptation

As shown in Figure 3, the process involves four stages: model and datasets preparation, template encapsulation, computing scheduling, and answer generation.

3.2.1 Model and Dataset Preparation

This phase encompasses the establishment of model and dataset configurations for the ensuing stages. Users have the option to employ their own dataset and model configurations to expand the current benchmark or to evaluate their own models using MathEval. For the model configuration, the elements include:

(1) Name: the identifier for the model being used;

(2) Prompt template: the general structure of prompts used by the model;

(3) System prompt (MSP): the official system prompt from the model description or technical report;

(4) User prompt (MUP): a token or phrase indicating the start or end of a user message;

(5) Bot prompt (MBP): similar to the user prompt, indicates the start or end of a bot response.

For the dataset configuration, the components consist of:

(1) Data metadata: information used to populate different parts of the template;

(2) Question prompt (DQP): Indicates where the question is located and specifies the different types of questions;

(3) Answer prompt (DAP): Specifies the kind of answer that needs to be generated, option from A to D or a specific answer;

(4) Options prompt (DOP): Indicates where the options are located within the template;

(5) CoT prompt: Guides the model to output different types of CoT reasoning for each dataset. For example, in multiple-choice questions, the CoT should reason through each choice before providing the final answer.

These detailed preparation ensure that both the model and datasets are configured correctly to facilitate accurate and contextually appropriate responses in the following stages. More details are discussed in the Electronic Supplementary Material.

3.2.2 Template Encapsulation

We encapsulate our final input prompt based on both model and dataset configurations. There are two scenarios: zero-shot prompt and few-shot prompt. Both settings use a combination of the previously discussed configuration elements. We include these two scenarios because base LLM models are generally not proficient in zero-shot scenarios, as they tend to continue generating content beyond the desired response. Introducing few-shot examples allows for a fair comparison by providing context and examples, thereby guiding the model to generate more accurate and contextually appropriate answers.

3.2.3 Calculation Scheduling and Answer Generation

The final two stages of our methodology are calculation scheduling and answer generation. In the calculation scheduling stage, the task is automatically partitioned

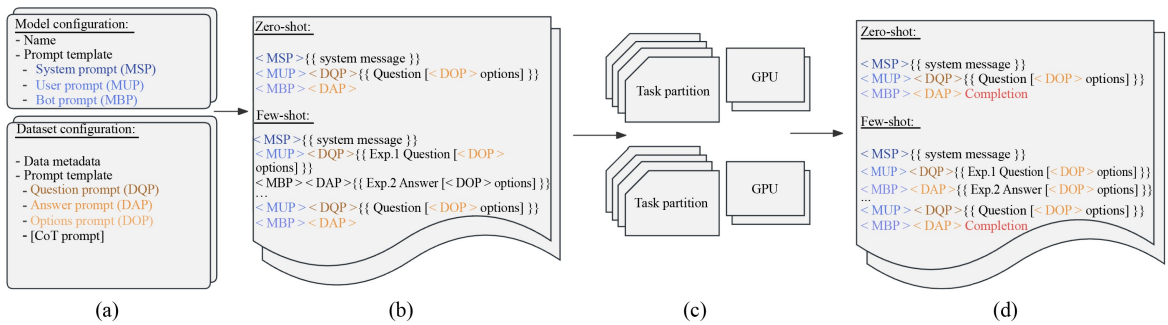

Figure 3 Four stages of the answer generation process. (a) Model and datasets preparations; (b) template encapsulation; (c) calculation scheduling; (d) answer generation. CoT: chain-of-thought.

according to available computational resources. This partitioning also takes into account the model size and dataset size, ensuring an efficient parallel processing to expedite the inference stage. Ultimately, this results in the generation of completion answers for both zero-shot and few-shot scenarios.

3.3 | Evaluation Methods

The conventional metric for evaluation entails designing specific answer extraction rules tailored to the models and datasets, followed by a matching process. While this traditional approach can yield stable results, it suffers from a lack of robustness. Minor variations in the output can lead to significantly different outcomes. Moreover, crafting answer extraction rules for each model and dataset based on their output formats introduces quadratic complexity, making rule-based evaluation criteria inefficient. Consequently, we adopt a general evaluation method that can be easily and cost-effectively extended to new datasets and models, thereby enhancing the fairness of the evaluation process through a unified standard.

We initially employed a two-stage evaluation method, as depicted in Figures 4(a) and 4(b). In first stage (Figure 4(a)), the generated response is processed to isolate the specific answer. In the subsequent stage (Figure 4(b)), the extracted answer is compared against the ground truth to produce a comparison result. Importantly, GPT-4 does not alter the underlying dataset in any way. Rather, it is leveraged solely to judge the correctness of the generated answers.

Given the robustness of methods based on LLMs, such as GPT-4 (OpenAI et al., 2023), these models exhibit strong comprehension capabilities and can handle diverse output formats. In contrast, rule-based methods offer greater stability in obtaining results. Consequently, in both evaluation pipelines, we primarily utilize outputs from GPT-4, supplementing them with regex-based results when GPT-4 fails. Detailed instructions for answer extraction and verification using GPT-4 and a comprehensive comparison between the regex-rule-based method and our GPT-4-as-judgement method are provided in the Electronic Supplementary Material.

Subsequently, as illustrated in Figures 4(c) and 4(d), we developed an answer comparison model that took as input the question, the model-generated answer, and the reference (golden) answer, and output a detailed, step-by-step analysis to extract and assess the correctness of the generated answers. This approach provides a comprehensive, end-to-end evaluation framework by unifying what was previously a bifurcated process, thereby enhancing both stability and cost-efficiency. An example of the training data used is provided in the Electronic Supplementary Material. Specifically, we employed 2,217,328 such evaluation instances derived from the GPT-4's two-stage evaluation process to train our DeepSeek-7B-based answer comparison model (Shao et al., 2024). Although GPT-4 furnished the evaluation judgments, it did not alter the dataset. Rather, its predictions served as training signals that enabled our model to emulate GPT-4's verification process. This integrated evaluation structure facilitates a more robust and cost-effective method for assessing model-generated answers. Structures allowed for a more stable and cost-efficient approach to evaluating generated answers.

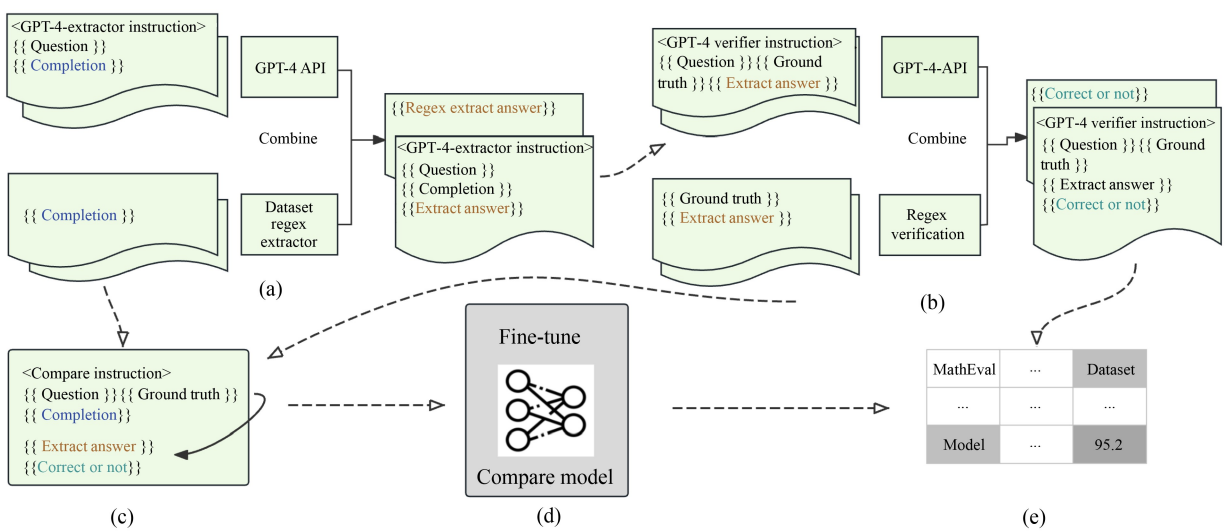

Figure 4 Two evaluation methods. (a) Answer extraction; (b) answer verification; (c) training data construction; (d) compard model training; (e) leaderboard refresh. (a), (b), and (e) depict a two-stage method involving answer extraction and verification using GPT-4; (c), (d), and (e) illustrate the training data construction of the comparison model and its training process.

To further foster transparency and reproducibility, we provide comprehensive details on how the compare-answer model is trained. Specifically, the model is optimized under standard log-likelihood objectives as a language model, in which both the reasoning chain and the final correctness judgment serve as training targets. Importantly, the input portion of each sample remains masked and does not contribute to the loss, thereby emphasizing the model's capability to infer correct judgments from its internally generated reasoning sequence. We do not perform data augmentation beyond minimal preprocessing steps (e.g., ensuring uniform formatting). While both our principal two-stage evaluation (using GPT-4) and this integrated model-based evaluation have distinct advantages, we primarily rely on the former for final reporting.

4 Datasets Details

4.1 | Unique Dataset Description

Arith3K is a high quality arithmetic evaluation dataset we constructed, consists of 3 main categories and 15 sub-categories, totaling 3,000 problems. It includes 12 types of mathematical operations, ranging from simple arithmetic and logarithmic operations to factorials, trigonometric functions, and complex compound operations. We systematically combined numbers and operators, and used Python code along with SymPy to verify the correctness of each expression. This makes Arith3K the most challenging dataset among arithmetic collections in our benchmark, designed to comprehensively assess the computational abilities of LLMs across varied difficulty levels.

TAL-SCQ5K-EN and TAL-SCQ5K-CN are comprehensive mathematical competition datasets available in English and Chinese, respectively. Each dataset comprises 5,000 multiple-choice questions, divided into 3,000 for training and 2,000 for testing, covering primary, middle, and high school level mathematics. These datasets are particularly valuable for CoT training as they include detailed solution steps. Furthermore, all mathematical expressions within the questions are formatted in standard text-mode LaTeX, ensuring clarity and consistency in presentation. To maintain the high quality of the TAL-SCQ datasets, each question undergoes a rigorous review process by two qualified teachers before being included. Moreover, an independent quality validation is conducted on a randomly selected sample of 200 problems, all of which are approved by independent teachers, with no identified issues. The Gaokao-2023 and Gaokao-2024 datasets are derived from the most recent Chinese National College Entrance Examination and consist of both multiple-choice and MWPs. These datasets, which reflect actual exam content, will be updated on an annual basis with forthcoming versions such as Gaokao-2025. These consistent updates are designed to help alleviate potential contamination of test data.

We focus on K-12 education levels due to their broad applicability and the availability of extensive datasets. However, we recognize that incorporating higher-level mathematics such as undergraduate topics and competition math problems like PutnamBench would provide deeper insights into the models' capabilities across varying difficulty levels. We are actively working to include these more challenging problems in future iterations of MathEval.

4.2 | Dataset Categorization and Detailed Problem Analysis

In this section, we conduct a detailed analysis of the differences among the datasets and classify them to avoid the issue of measuring the same ability dimensions. First, in Table 1, we present each of our datasets and their corresponding classifications, including language, problem type, and corresponding grade level (which can partially reflect the difficulty level).

To further analyze the distinctions between the datasets, we examined the data distribution of problems in the different datasets to ensure that they are dissimilar. For each dataset, we first randomly sampled 200 query problems and obtained their representations using the Llama-3-8B (Grattafiori et al., 2024) model. We performed t-SNE dimension reduction on these representations, with the cosine similarity situation in Figure 5 and the visualization shown in Figure 6.

We computed the cosine similarity between each query across every pair of datasets and finally took the absolute value of the average as their correlation. We found that the correlation scale between dataset queries ranged from 0 to 0.8, and according to statistics, 75.32% were less than 0.6, and 60.17% were less than 0.5. This demonstrates the dissimilarity between dataset queries, reflecting that to some extent they measure different abilities.

Moreover, we made a surprising discovery in the *t*-SNE results: The *t*-SNE naturally formed three clusters. In Figure 6, the cluster on the upper-left consists of English datasets, the cluster on the upper-right consists of Chinese datasets, and the cluster at the bottom consists of arithmetic problems. From component 1, we can observe that on the left are the English datasets, in the middle are the arithmetic problems, and on the right are the Chinese datasets.

Further observing the Chinese and English clusters, we found that as the *t*-SNE component 2 value

Table 1 Twenty-two datasets used in MathEval with references and three-dimensional categories

Dataset	Language	Type	Grade	Number
AGIEval-Math (Zhong et al., 2024)	Chinese	MWP	H	1,943
Ape210K (Zhao et al., 2020)	Chinese	MWP	P	5,000
Arith3K (Ours)		Arithmetic	M	3,000
ASDiv-A (Miao et al., 2020)	English	MWP	P	122
BBH-Math (Suzgun et al., 2023)		Arithmetic	P	250
CMMLU-Math (Li et al., 2023a)	Chinese	MWP	H	499
Dolphin1878 (Shi et al., 2015)	English	MWP	P	187
Gaokao-2023 (Ours)	Chinese	MWP	H	156
Gaokao-2024 (Ours)	Chinese	MWP	H	216
Gaokao-Math (Zhang et al., 2023)	Chinese	MWP	H	432
GSM8K (Cobbe et al., 2021)	English	MWP	P	1,319
MAWPS (Koncel-Kedziorski et al., 2016)	English	MWP	P	238
Math23K (Wang et al., 2017)	Chinese	MWP	P	1,000
Math401 (Yuan et al., 2023)		Arithmetic	P	401
MATH (Hendrycks et al., 2021a)	English	MWP	H	5,000
MathQA (Amini et al., 2019)	English	MWP	P	2,985
MMLU-Math (Hendrycks et al., 2021b)	English	MWP	H	848
TAL-SCQ5K-CN (Ours)	Chinese	MWP	P	2,000
TAL-SCQ5K-EN (Ours)	English	MWP	P	1,998
SVAMP (Patel et al., 2021)	English	MWP	P	1,000
OlympiadBench-CN (He et al., 2024)	Chinese	MWP	H	675
OlympiadBench-EN (He et al., 2024)	English	MWP	H	675

Notes. H: high school; M: middle school; P: primary school; MWP: math word problem; BBH: Big-Bench-Hard.

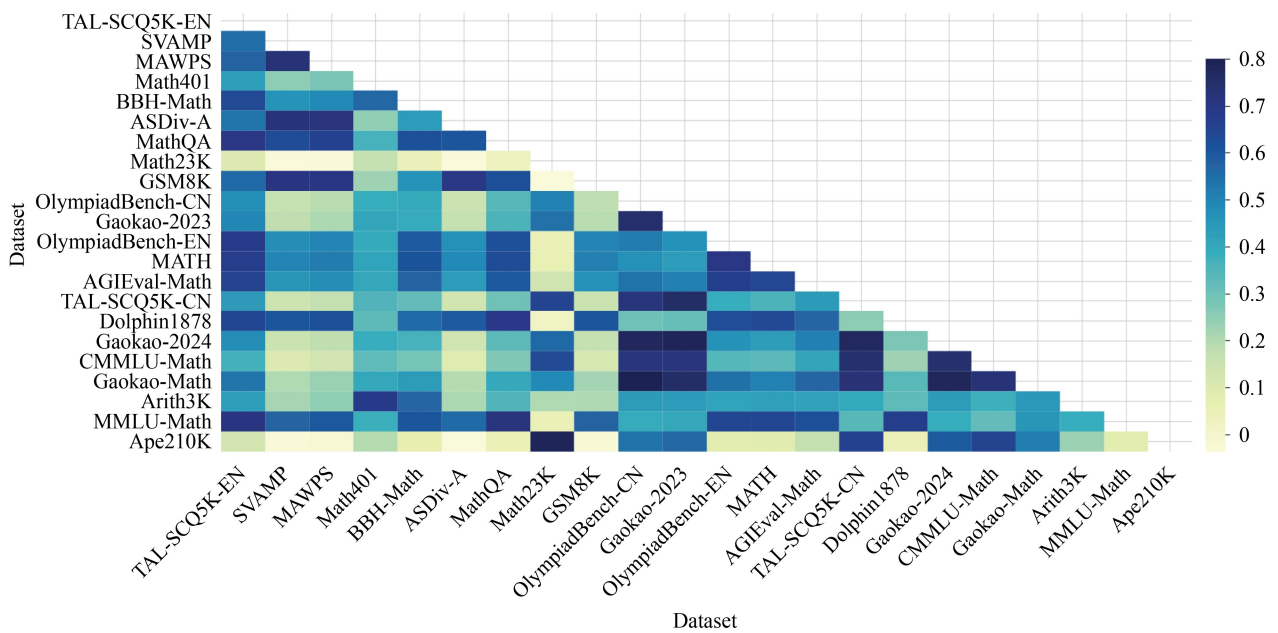

Figure 5 Average cosine similarity between datasets. BBH: Big-Bench-Hard.

decreased (from top to bottom in the figure), the problems became increasingly difficult, and the corresponding grade levels also rose. This perfectly reflects that our difficulty levels are distributed across

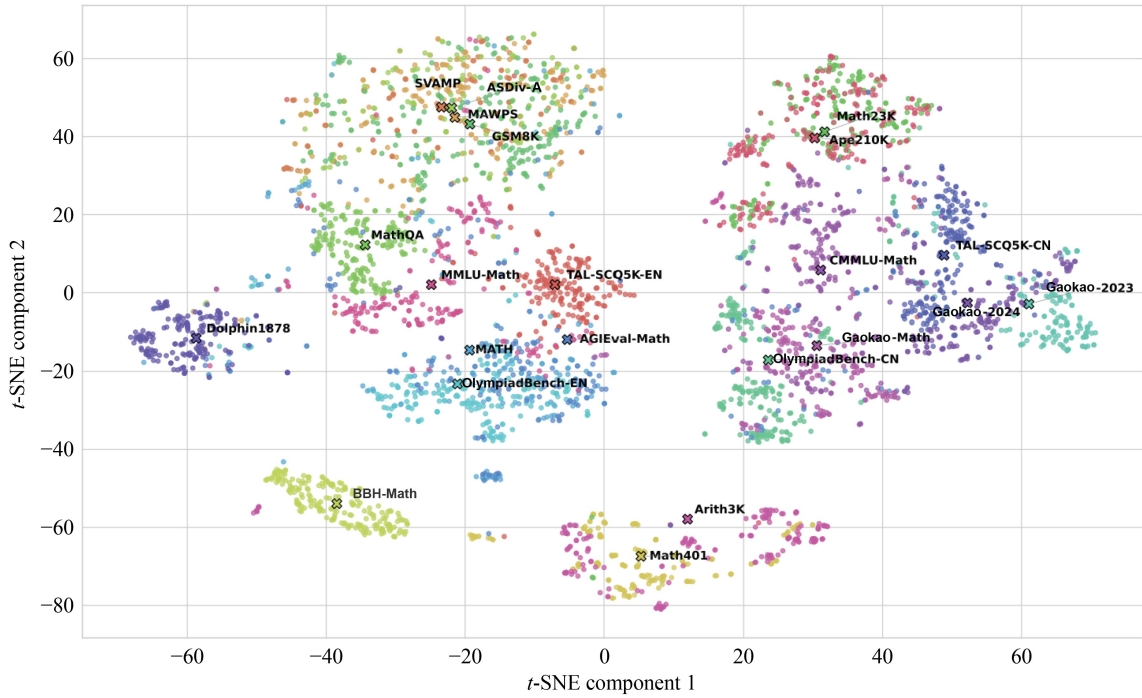

Figure 6 *t*-SNE visualization of query embeddings. BBH: Big-Bench-Hard.

various grades, with corresponding datasets at each level. This shows that our 22 datasets measure different aspects, whether in terms of difficulty or language.

5 Experiment

5.1 | Evaluated Models

Fifty-two different models have been evaluated, we have categorized these models into three distinct groups. The first group consists of open-source models, characterized by their accessibility in terms of model weights and architecture. The second group comprises closed-source models, which are accessible only through APIs without disclosure of their underlying architecture or weights. The third group specifically includes open-source models that have been fine-tuned on math domain data, allowing for a more tailored analysis in this specific area. Models within each categorized group are listed in the electronic supplementary material A.

5.2 | Compare Answer Models

In this section, we explore the methodologies selected for the compare-answer task, a notably challenging aspect of mathematical benchmark. This complexity is well-acknowledged within the academic community, as highlighted in several case studies included in the Electronic Supplementary Material. Many benchmarks in this field also encounter difficulties due to their reliance on rule-based extraction and matching approaches. These methods typically struggle to accommodate the diverse output behaviors exhibited by different models. Furthermore, there has been a lack of focused research on this domain, particularly in terms of a systematic analysis of the accuracy of answer matching across various models. This section aims to address these gaps by detailing our approach and the rationale behind our methodological choices.

To authenticate the precision of our methodology, we have organized a large-scale human annotation process of the model output results, which was carried out over the course of approximately one month. We have selected GPT-4, DeepSeek-Math-7B-Base, DeepSeek-Math-7B-Instruct, and DeepSeek-Math-7B-RL as the basis for validation. Five annotators were assigned to annotate each line of outputs, and the majority vote result was considered the final decision on the correctness of the model output. The details of the data can be found in the Electronic Supplementary Material. The overall Fleiss' Kappa (Fleiss, 1971) achieved a score of 0.8871, indicating significant inter-annotator agreement. We believe the human annotation result is reliable and treats the overall average accuracy as the golden standard.

We evaluated two methods for answer comparison. The first is our two-stage GPT-4-based judgment depicted in Subsection 3.3. The second is a fine-tuned DeepSeek-7B-Base model (Fine-tuned DeepSeek) trained on our private answer-comparison

data and GPT-4 output comparison data (partially verified by human annotators to fix potential errors). We computed the overall average accuracy for each answer-comparison method using 19 selected datasets out of 22 from the four chosen models. The results are shown in Figure 7, where the *y*-axis represents the absolute difference between the proposed answer-comparison method and the human evaluation result, with larger values indicating worse performance. We presented the zero-shot, few-shot, and best-of-two previous settings in this figure, focusing primarily on the best-of-two setting. Initially, we observed that GPT-4 performed consistently well across all models, with an absolute difference ranging from 0 to 0.1. Both methods performed poorly on the output of DeepSeek-Math-7B-Base, likely due to the base model's tendency to output useless tokens and inability to stop at the correct position, which poses challenges for the answer-comparison model. Notably, Fine-tuned DeepSeek achieved the same performance as human annotators on the output of DeepSeek-Math-7B-RL, demonstrating the effectiveness of our method. Given GPT-4's consistent performance, it will be our primary model for further analysis. We have open-sourced our custom Fine-tuned DeepSeek model to provide a viable alternative for users who lack access to GPT-4.

5.3 | Evaluation Results

The main results of MathEval are shown in Table 2, with detailed results for more models provided in the Electronic Supplementary Material. We calculated the average accuracy of each model across 22 datasets and ranked the models accordingly. The accuracy for each dataset and model is the percentage of correctly solved problems, and then the average value across the 22 datasets is used as the overall accuracy. In the subsequent analysis, the overall average is predominantly used as the primary metric for evaluation. As shown in Table 1, we categorized each dataset into different categories and reported the accuracy for each category in Table 2. For example, En (English) denotes the average accuracy for datasets categorized under the English language. To ensure the credibility of our evaluation results, as detailed in the Electronic Supplementary Material, we compared our evaluated results from GPT-4 with the reported metrics of each published model on the GSM8K and MATH datasets, which are commonly used for assessing math-solving abilities. The minor discrepancies observed demonstrate the reliability of our evaluation pipeline.

As we can see in Table 2, Claude-3.5-Sonnet, a closed-source model, has demonstrated exceptional performance, surpassing GPT-4 by a significant margin with an average accuracy of 77.0%. This superiority is evident across various dimensions, particularly in its advanced understanding of both English and Chinese languages. Claude-3.5-Sonnet's proficiency in handling high school level problems further highlights its reasoning capabilities.

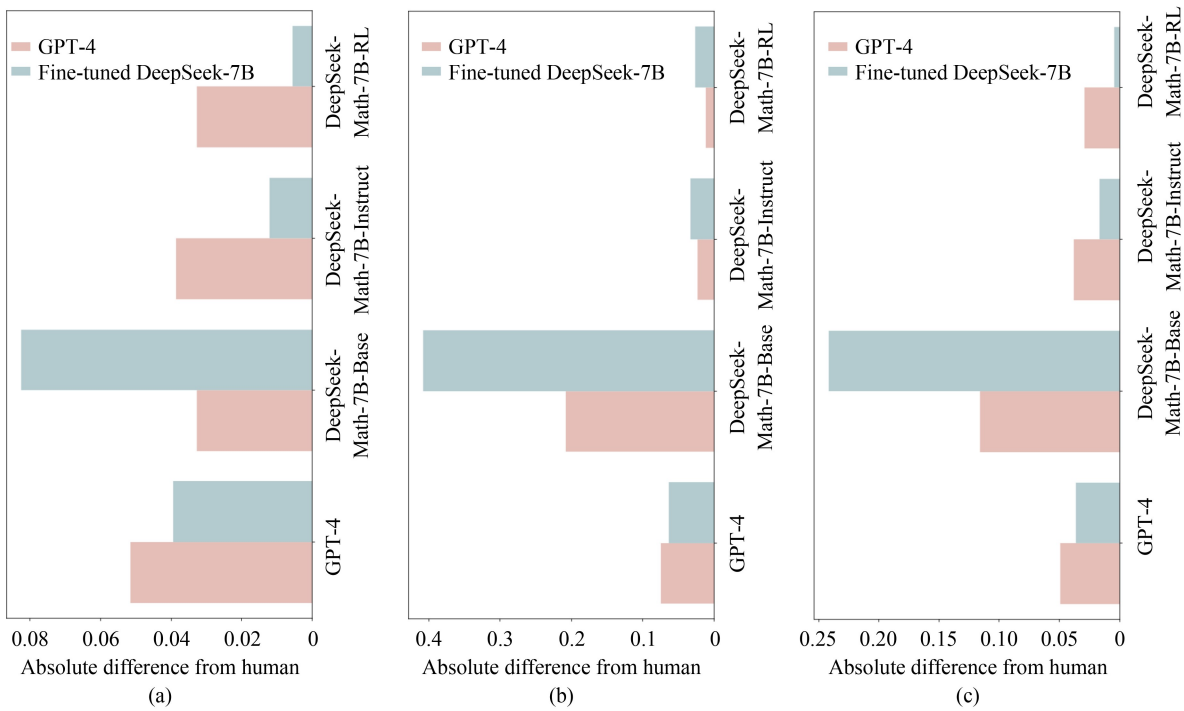

Figure 7 Comparison of absolute differences from human answers across different compare-answer models. (a) Zero-shot; (b) few-shot; (c) best of two settings.

Table 2 Summary of principal outcomes from MathEval

Category	Models	Language		Type		Grade			Avg(%)
		En (%)	Cn (%)	MWP (%)	Arith (%)	Prim (%)	Mid (%)	High (%)	
Closed-source models	Claude-3.5-Sonnet	84.7	67.2	76.4	80.8	90.0	57.3	61.8	77.0
	WenXin 4.0	78.3	65.7	72.4	93.1	88.2	89.6	56.3	75.2
	Gemini 1.5 Pro	81.9	63.8	73.3	81.9	88.8	46.9	58.5	74.5
	GLM4	76.5	61.3	69.3	60.9	83.1	32.4	52.2	68.1
	Spark-3.5	72.8	60.6	67.0	68.4	81.5	41.2	51.1	67.2
	GPT-4	72.4	45.9	59.8	67.1	79.6	38.3	38.3	60.8
Open-source models	Qwen2-72B-Instruct	81.8	64.7	73.7	78.7	88.7	57.3	57.2	74.4
	Qwen1.5-110B-Chat	76.3	57.3	67.3	68.6	84.0	40.8	48.4	67.5
	Qwen2-72B	73.0	57.0	65.4	65.4	79.7	35.3	49.7	65.4
	Llama-3-70B-Instruct	76.6	51.7	64.8	68.8	82.3	42.4	45.3	65.4
	Qwen2-7B-Instruct	75.8	52.7	64.8	67.4	81.3	46.3	45.8	65.2
	Qwen1.5-72B-Chat	71.7	55.1	63.8	62.8	79.6	33.4	45.7	63.7
Math domain models	DeepSeek-Math-7B-RL	74.0	50.3	62.8	64.4	79.5	44.0	43.0	63.0
	DeepSeek-Math-7B-Instruct	69.7	46.7	58.8	57.7	75.7	36.6	38.3	58.7
	InternLM2-Math-20B	66.0	44.7	55.9	41.3	68.4	28.8	37.4	53.9
	MetaMath-70B	57.6	27.7	43.4	32.1	58.3	12.5	23.3	41.9
	MAmmoTH-70B	56.5	27.6	42.8	30.9	56.6	11.4	23.9	41.2
	GAIRMath-Abel-70B	53.5	30.8	42.7	25.5	53.3	11.5	26.3	40.4

Notes. En: English, Cn: Chinese, Arith: arithmetic, Prim: primary school, Mid: middle school, High: high school, Avg: average score. The table displays the top-six performing models in each category.

For open-source models, Qwen2-72B-Instruct leads the pack with an impressive average accuracy of 74.4%. This model's performance is followed closely by Qwen15-110B-Chat, which also shows strong results, indicating that the newer large-parameter chat models possess superior mathematical abilities.

In the math domain, despite having only 7 billion parameters, DeepSeek-Math-7B-RL stands out with an average accuracy of 63.0%, showcasing the effectiveness of its post-training. However, it still trails behind the top-5 open-source models, which are all outperformed by the top-3 closed-source models. This underscores the importance of model parameter size in achieving leading mathematical capabilities and highlights the current gap between open-source and closed-source models.

5.4 | Discussion

With MathEval, we have uncovered several intriguing insights. We will delve into these findings in detail within this section.

5.4.1 Closed-Source Models

Closed-source models exhibit a higher capability range than open-source models and math domain models. As shown in the Figure 8(a), not only does it exhibit the highest capability ceiling, but it also maintains a high capability floor, with only GPT-3.5 lagging slightly. This indicates that closed-source models typically exhibit consistently superior performance in mathematical tasks. Nevertheless, we also observed that the 25th percentile range of closed-source models is encompassed by the capability range of open-source models. This suggests that excellent open-source models can achieve performance comparable to closed-source models.

5.4.2 Open-Source Models

Open-source models exhibit a wide range of capabilities influenced by both the type of base model and the size of the model parameters, as shown in Figure 8(b). While the size of the model parameters does not directly determine the model's mathematical abilities, it can increase the potential upper limit of these abilities. Consistent with general conclusions, we observe that the mathematical ability of models with the same base architecture has a linear relationship with the logarithm of their parameter sizes. Additionally, chat models consistently outperform base models, reflecting the stabilizing effect of post-training. Furthermore, analyzing the lines of similar color in the figure reveals

that the slopes of models with the same base architecture are remarkably uniform. Interestingly, newer series exhibit steeper slopes, indicating that their mathematical abilities improve more effectively with an increase in parameter size.

5.4.3 Problem Type Dimension

For problem type dimension, as shown in Figure 8(d), the scarcity of arithmetic-related datasets leads to significant fluctuations in arithmetic capabilities across models, represented by the blue line. Models highlighted in blue, positioned below the average difference line, exhibit stronger arithmetic abilities compared to their capabilities in solving MWPs. For closed-source models, the notable deviations of WenXin 4.0 may be due to their arithmetic plugins. We did not use API versions with plugins for GPT-4, which could affect their performance in arithmetic tasks. Other open-source models like Llama-3-70B and InternLM2-20B-Base also show strong arithmetic capabilities. Conversely, models above the average difference line, highlighted in red, are predominantly fine-tuned on MWPs data. This includes specialized models such as MAmmoTH-70B, MetaMath-70B, and

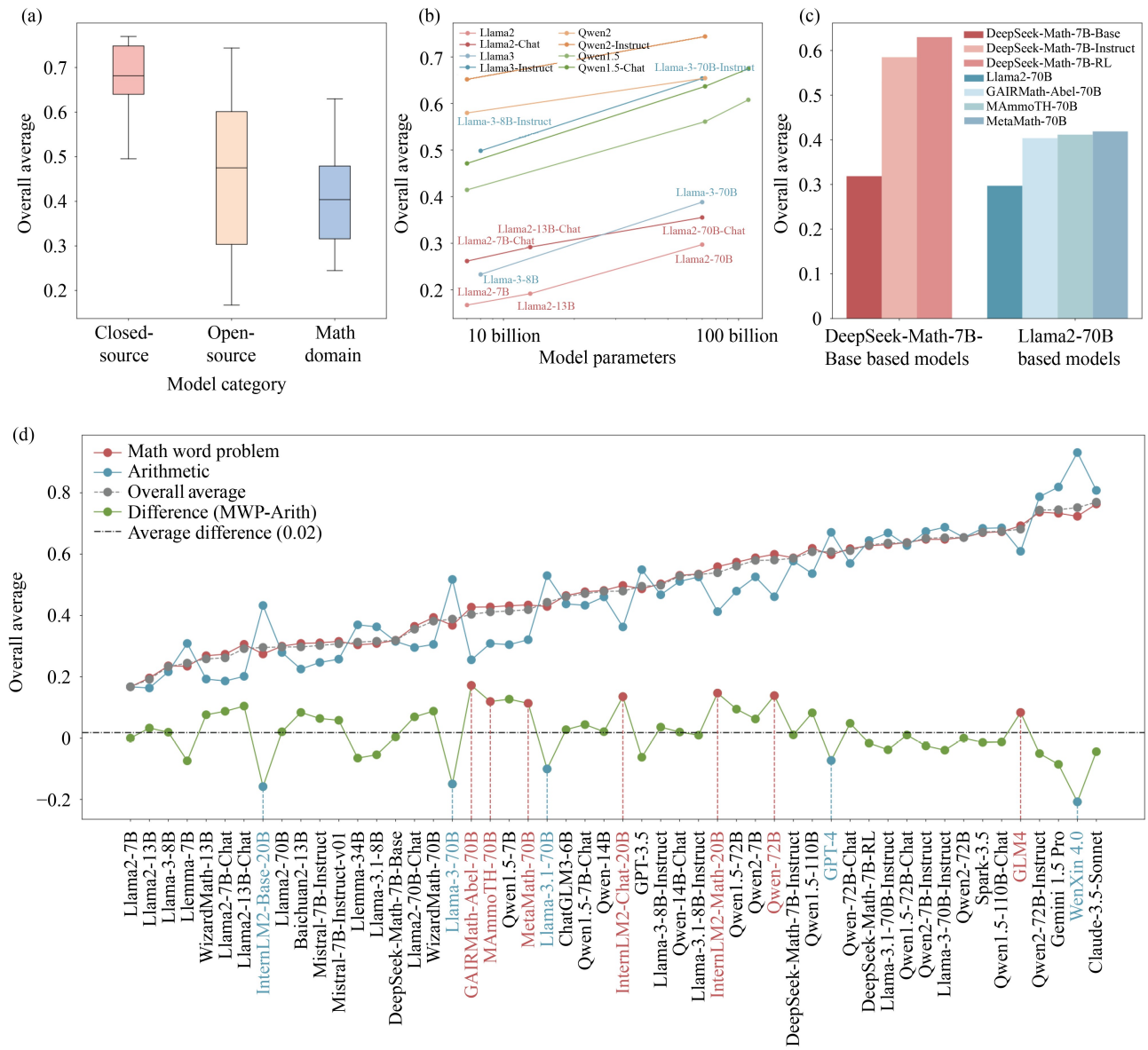

Figure 8 MathEval evaluation results. (a) Overall average for different model categories; (b) model performance by parameter size; (c) improvements on math domain models; (d) comparison of solving arithmetic and math word problems capabilities of models. (a), (b), and (c) Show the discovery of closed-source models, open-source models, and math domain models. (d) Compares the model-level capabilities across problem type dimensions.

GAIRMath-Abel-70b. A similar trend is evident in InternLM2-Math-20B and InternLM2-Chat-20B, which, unlike the more arithmetic-proficient InternLM2-20B-Base, likely benefited from targeted fine-tuning on MWPs datasets during the post-training phase. Additionally, models like Qwen1.5-72B, Qwen-72B, and GLM4 also demonstrate enhanced capabilities in handling MWPs.

5.4.4 Few-Shot/Zero-Shot Setting

Few-shot/zero-shot setting is a relatively consistent part of prompts adaptation. We aim to maintain consistency while ensuring fairness, making it important to understand how different few-shot/zero-shot settings affect model capabilities. Specifically, our evaluation includes three settings: few-shot, zero-shot, and the higher accuracy between few-shot and zero-shot at the dataset level. As shown in the Figure 9, the "dataset-level higher" setting consistently outperforms using either few-shot or zero-shot alone across all models. It also produces smoother curves with fewer outliers, indicating that this setting contributes to the robustness and fairness of the evaluation. When comparing zero-shot and few-shot, zero-shot generally performs better on most models, with only some base models showing significantly lower performance (represented by red dashed lines). Notably, the base models in the Qwen series do not exhibit this phenomenon.

5.4.5 Language Dimension

For the language dimension, as illustrated in Figure 10(a), we observed that the mathematical capabilities in Chinese consistently trail those in English. To account for the potential impact of problem difficulty, we compared the math capabilities in English and Chinese separately for primary and high school grade subsets. As Figures 10(b) and 10(c) demonstrate, this trend persists in the primary school subset, while the differences between Chinese and English capabilities are negligible in the high school subset. This could be attributed to primary school problems requiring more language comprehension. Models with stronger Chinese mathematical abilities, such as WenXin 4.0 and Spark-3.5, primarily developed by Chinese companies, are displayed in blue font below the average difference line. Conversely, models with stronger English capabilities, represented in red, include Mistral-7B-Instruct, Llama2-70B-Chat, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct, are instruction fine-tuned models primarily developed by companies based in English-speaking countries, and may exhibit weaker performance in Chinese due to the relative scarcity of Chinese math problems in their fine-tuning data. Another category, including MAmmoTH-70B and MetaMath-70B, comprises math domain fine-tuned models that exclusively use augmentation data from English datasets.

5.4.6 Math Domain Models

Math domain models enhance the capabilities of base models by leveraging specialized data from the mathematical domain for continued pre-training,

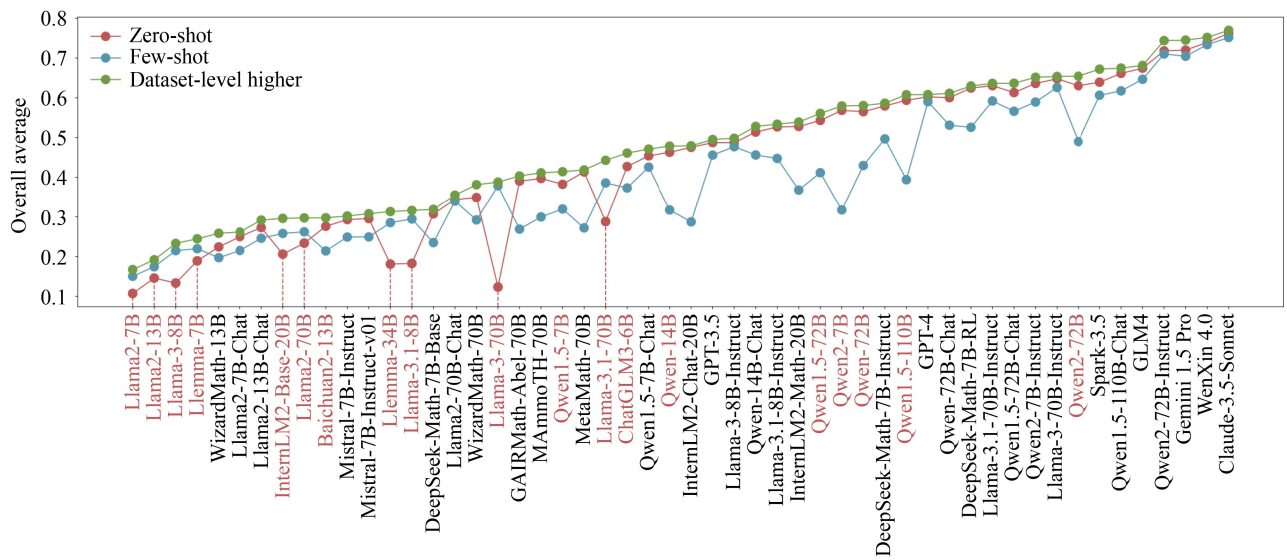

Figure 9 Comparison of prompts adaptation settings among few-shot, zero-shot, and dataset-level higher. Base models are highlighted in red on the *x*-axis, while post-training models are shown in black.

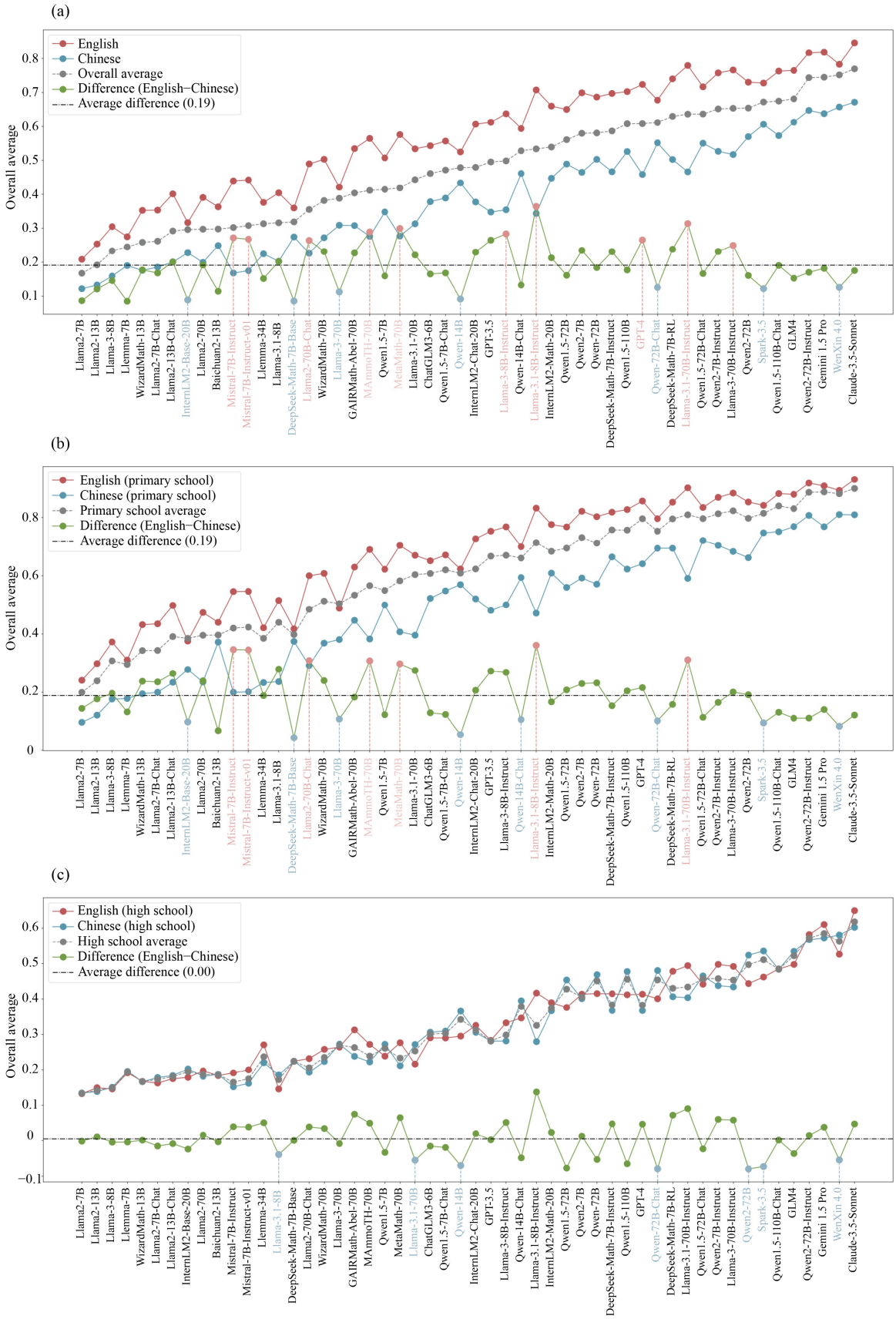

Figure 10 Comparison of math capabilities between Chinese and English language. (a) In all MWP datasets; (b) in primary school subsets; (c) in high school subsets. MWP: math word problem.

supervised fine-tuning and reinforcement learning. As shown in Figure 8(c), models fine-tuned on Llama2-70B and DeepSeek-Math-7B-Base exhibit more than double the improvement, highlighting that post-training significantly boosts the model's specialized abilities, extending beyond specific datasets.

5.4.7 Comparison of Mathematical Abilities Across Three Dimensions

As shown in Figures 8(d), 10, and 11, models tend to exhibit consistent performance within the same dimension, such as language, grade, or problem type. For example, a model that performs well on English problems is likely to perform similarly on Chinese problems. However, evaluating different types of mathematical abilities is crucial not only for completeness but also to identify relative differences in model capabilities. These differences, often resulting from the model's data and training process, provide valuable insights for future improvements.

5.4.8 Grade Dimension

Given the presence of only one middle school dataset, our discussion will center on the model capabilities for primary and high school math problems. As Figure 11(a) illustrates, models consistently perform better on primary school problems than on high school problems, likely due to inherent differences in difficulty. Notably, Claude-3.5-Sonnet and Gemini 1.5 Pro, demonstrate significantly higher accuracy on primary school problems. This may be attributed to the stronger comprehension abilities of these models, as primary school problems are predominantly word problems. Conversely, the Llemma-7B and Llemma-34B models, display a smaller advantage. We hypothesized that this could be due to their pre-training data, created with AlgebraicStack, which contained complex mathematical knowledge, including symbolic and formal math. Additionally, in Figures 11(b) and 11(c), we re-evaluated the models' capabilities based on problem difficulty within the Chinese and English subsets. We found that only GPT-3.5 showed a weakened strength in primary school math abilities within the Chinese subset. The other conclusions remain largely consistent.

5.4.9 Potential Data Contamination

By conducting comprehensive evaluations across all datasets, we identified potential data contamination issues that were not apparent when analyzing a small subset of data. Specifically, Figure 12 illustrates discrepancies in model performance on the Gaokao-2023 dataset—A newly introduced set of questions that none of the models had encountered during training or fine-tuning phases. In the upper chart of Figure 12, we presented the Chinese subsets rank (blue bars) and the Gaokao-2023 rank changes (orange and green bars) for each model. A smaller rank indicates better performance. The orange bars represent models whose rank increased (indicating poorer performance) on Gaokao-2023 relative to other datasets, while the green bars represent models whose rank decreased (indicating better performance) on Gaokao-2023. Our analysis reveals that certain models, notably ChatGLM3-6B and Baichuan2-13B, exhibit a significant increase in rank when evaluated on Gaokao-2023, suggesting a drop in their relative performance on this new dataset. This discrepancy implies that these models may have benefited from potential data contamination in the other datasets, artificially inflating their performance. Furthermore, many of the Qwen-series models display orange bars, indicating a deterioration in their performance ranking on Gaokao-2023 compared to other datasets. This pattern suggests that these models may have been trained on data overlapping with our evaluation sets, leading to inflated performance on those datasets but not on the unseen Gaokao-2023. In contrast, most base models (those not undergoing supervised fine-turning and reinforcement learning from human feedback) exhibit green bars, improving their performance ranking on Gaokao-2023. This observation supports the notion that chat models are more susceptible to data contamination due to their exposure to a wider range of data during instruction finetuning stages, which may include similar mathematical word problems.

6 Limitations and Future Directions

Despite the progress made in developing our MathEval framework and evaluation protocols, several limitations and future works must be acknowledged.

6.1 | Limited Coverage of Middle School-Level MWP

Currently, the benchmark contains relatively few MWPs explicitly designed at the middle school level. This limitation may reduce its effectiveness in assessing foundational reasoning skills and the gradual progression of difficulty between primary- and high-school-level tasks. To address this, future versions will include a more balanced range of middle-school word problems. By focusing explicitly on these transitional levels, we aim to ensure a smoother difficulty progression and better capture foundational reasoning skills.

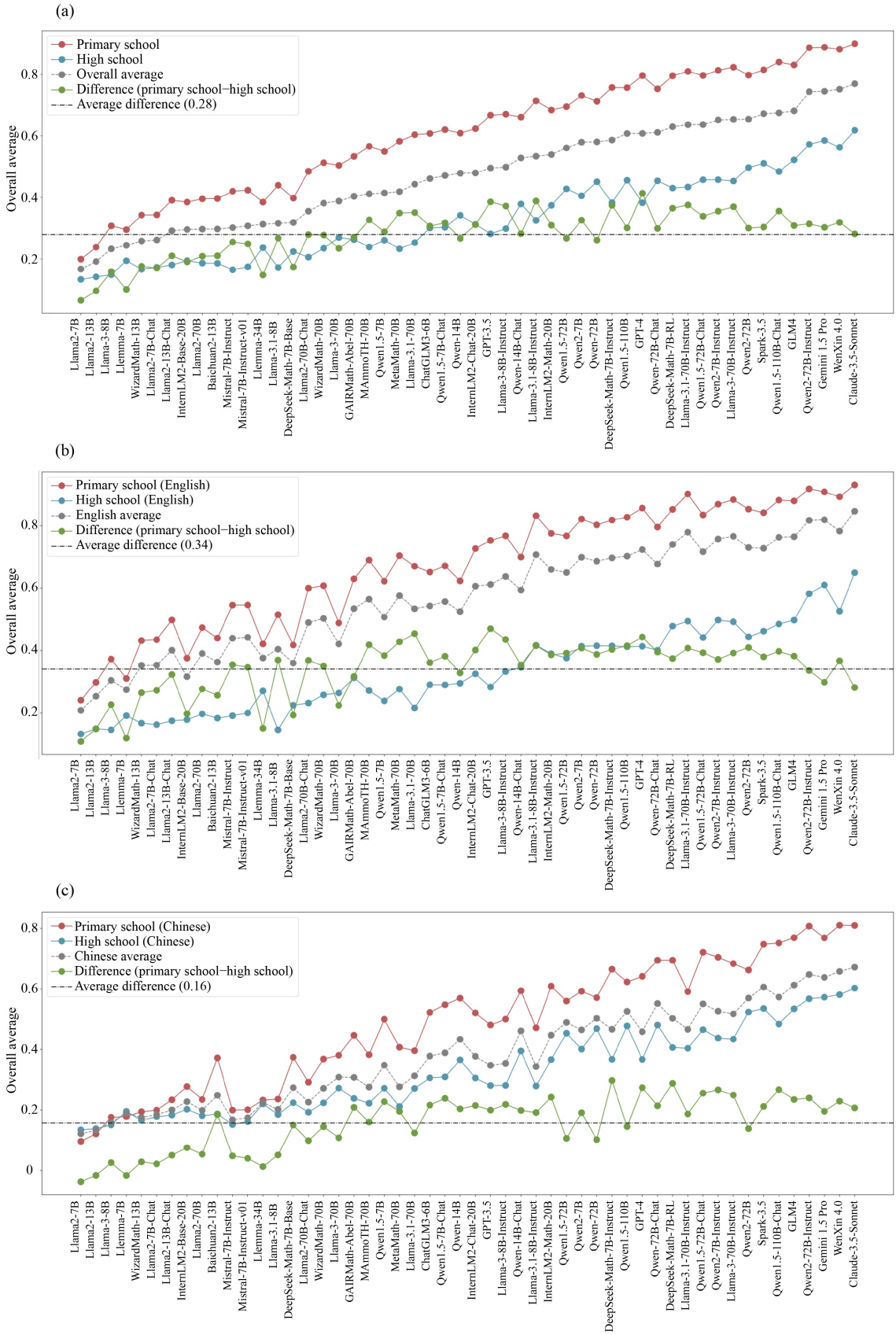

Figure 11 Comparison of math capabilities between primary and high school. (a) In all MWP datasets; (b) in English subsets; (c) in Chinese subsets. MWP: math word problem.

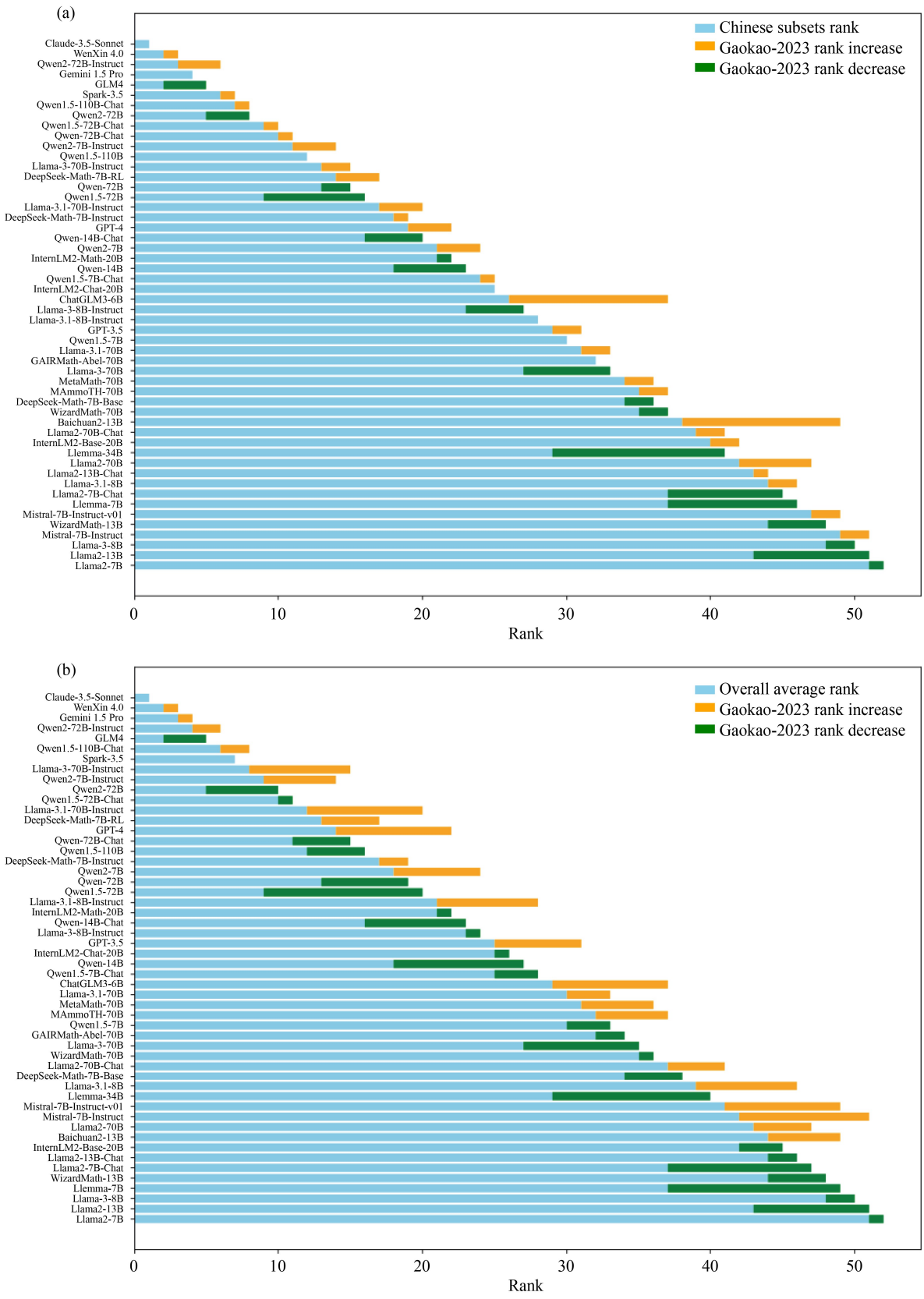

Figure 12 (a) Chinese subsets rank and Gaokao-2023 rank change by model; (b) overall average rank and Gaokao-2023 rank change by model.

6.2 | Absence of Multi-Modal Mathematical Evaluations

Our current evaluation pipeline does not address tasks requiring visual or diagrammatic reasoning, such as geometry problems involving figures or proofs that rely on visual cues. Future iterations will incorporate problems featuring diagrams, geometric shapes, and visual data representations (e.g., charts, graphs) alongside textual descriptions. This enhancement will allow us to evaluate models' abilities to interpret and reason about images in addition to text, thereby providing a more comprehensive assessment.

6.3 | Lack of Finer-Grained Difficulty Classification and Task-Specific Categorization

Although our dataset includes a range of difficulty levels and various types of problems, it lacks finer-grained difficulty labeling (e.g., based on the number of reasoning steps required) and specialized categorization (e.g., geometry, number theory, calculus). Implementing more granular difficulty labels and task-specific categories would enable more targeted benchmarking and provide deeper insights into specific areas where models excel or struggle. Addressing this gap is a key direction for our future work.

By acknowledging these limitations and outlining clear paths for improvement, we hope to enhance the robustness and comprehensiveness of the MathEval framework, ultimately advancing the field of mathematical problem-solving evaluation.

7 Conclusions

In this paper, we proposed MathEval, the first comprehensive evaluation benchmark for the mathematical capabilities of LLMs. Our evaluation encompassed 52 models across 22 datasets, organized into distinct scenarios along three dimensions. Our pipeline facilitates flexible adaptation to various datasets and models. Moreover, we propose an LLM-based approach for the automatic extraction and verification of mathematical answers, serving as a general and precise metric. We hope that MathEval will help provide an impartial evaluation of the mathematical abilities of LLMs, advancing the continuous improvement of LLM mathematical capabilities and expanding practical applications.

Acknowledgments This work was supported in part by the National Key R & D Program of China (Grant No. 2022YFC3303600), in part by the National Natural Science Foundation of China (Grant No. 62477025), in part by the Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (Grant No. 2022LSYS003), and in part by Beijing Municipal Science and Technology Project (Grant No. Z241100001324011).

Conflict of interest The authors declare that they have no conflict of interest related to the content of this paper.

Authors Contributions Tianqiao Liu: Handled the engineering implementation of MathEval, including the development of the compare-answer model and its training process to enable precise answer validation without requiring GPT-4 access. Zui Chen: Conducted the visualization analysis and contributed to the completion of key tables, providing meaningful insights into the performance of LLMs across various datasets. Zhensheng Fang: Managed the collection of datasets and assisted in compiling critical tables within MathEval, ensuring a diverse and comprehensive set of problems for evaluation. Weiqi Luo: Provided writing guidance and contributed to the addition of a difficulty dimension to MathEval, enhancing the framework's ability to assess problem complexity and model performance across varying levels of difficulty. Mi Tian: Offered strategic guidance on the engineering approach and provided insights into the partitioning of test models and datasets, ensuring a robust and fair evaluation process. Zitao Liu: Responsible for the overall framework design of MathEval, ensuring a comprehensive and cohesive structure for evaluating the mathematical reasoning capabilities of large language models. Additionally, provided meticulous writing guidance to enhance the clarity and coherence of the framework's presentation. All authors whose names appear on the submission made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; drafted the work or revised it critically for important intellectual content; approved the version to be published; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Data Availability Statements The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s44366-025-0053-z and is accessible for authorized users.

References

Ahn, J., Verma, R., Lou, R. Z., Liu, D., Zhang, R., & Yin, W. P. (2024). Large language models for mathematical reasoning: Progresses and challenges. In: *Proceedings of the 18th Conference of the European Chapter of the Association for*

Computational Linguistics: Student Research Workshop. Stroudsburg: ACL, 225–237.

Amini, A., Gabriel, S., Lin, S. C., Koncel-Kedziorski, R., Choi, Y., & Hajishirzi, H. (2019). MathQA: Towards interpretable math word problem solving with operation-based formalisms. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: ACL, 2357–2367.

Bai, Y. S., Lv, X., Zhang, J. J., Lyu, H. C., Tang, J. K., Huang, Z. D., Du, Z. X., Liu, X., Zeng, A. H., Hou, L., & et al. (2023b). LongBench: A bilingual, multitask benchmark for long context understanding. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: ACL, 3119–3137.

Chen, J. H., Liu, Z. T., Hou, M. L., Zhao, X. Y., & Luo, W. Q. (2024). Multi-turn classroom dialogue dataset: Assessing student performance from one-on-one conversations. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. New York: ACM, 5333–5337.

Chen, Z., Liu, T. Q., Tian, M., Tong, Q., Luo, W. Q., & Liu, Z. T. (2025). Advancing math reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. *arXiv Preprint*, arXiv:2501.14002.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., & et al. (2021). Training verifiers to solve math word problems. *arXiv Preprint*, arXiv:2110.14168.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., & et al. (2024). The Llama 3 herd of models. *arXiv Preprint*, arXiv:2407.21783.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

He, C. Q., Luo, R. J., Bai, Y. Z., Hu, S. D., Thai, Z., Shen, J. H., Hu, J. Y., Han, X., Huang, Y. J., Zhang, Y. X., & et al. (2024). OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: ACL, 3828–3850.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021a). Measuring massive multitask language understanding. *arXiv Preprint*, arXiv:2009.03300.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021b). Measuring mathematical problem solving with the math dataset. *arXiv Preprint*, arXiv:2103.03874.

Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., & Hajishirzi, H. (2016). MAWPS: A math word problem repository. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: ACL, 1152–1157.

Li, H. N., Zhang, Y. X., Koto, F., Yang, Y. F., Zhao, H., Gong, Y. Y., Duan, N., & Baldwin, T. (2023a). CMMLU: Measuring massive multitask language understanding in Chinese. In: *Proceedings of the Findings of the Association for Computational Linguistics*. Stroudsburg: ACL, 11260–11285.

Li, J. Y., Cheng, X. X., Zhao, X., Nie, J. Y., & Wen, J. R. (2023b). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 6449–6464.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y. A., Narayanan, D., Wu, Y. H., Kumar, A., & et al. (2023). Holistic evaluation of language models. *arXiv Preprint*, arXiv:2211.09110.

Liu, T. Q., Chen, Z., Liu, Z. T., Tian, M., & Luo, W. Q. (2024). Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. *arXiv Preprint*, arXiv:2211.09110.

Ma, Y. R., Chen, Z., Liu, T. Q., Tian, M., Liu, Z., Liu, Z. T., & Luo, W. Q. (2024). What are step-level reward models rewarding? Counterintuitive findings from MCTS-boosted mathematical reasoning. *arXiv Preprint*, arXiv:2412.15904.

Miao, S. Y., Liang, C. C., & Su, K. Y. (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*. Stroudsburg: ACL, 975–984.

Mishra, S., Finlayson, M., Lu, P., Tang, L., Welleck, S., Baral, C., Rajpurohit, T., Tafjord, O., Sabharwal, A., Clark, P., & Kalyan, A. (2022). LILA: A unified benchmark for mathematical reasoning. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 5807–5832.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & et al. (2023). GPT-4 technical report. *arXiv Preprint*, arXiv:2303:08774.

OpenCompass. (2025). *Opencompass*. Available from Opencompass in GitHub.

Patel, A., Bhattamishra, S., & Goyal, N. (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*. Stroudsburg: ACL, 2080–2094.

Shao, Z. H., Wang, P. Y., Zhu, Q. H., Xu, R. X., Song, J. X., Bi, X., Zhang, H. W., Zhang, M. C., Li, Y. K., Wu, Y., & Guo, D. Y. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv Preprint*, arXiv:2402.03300.

Shi, S. M., Wang, Y. H., Lin, C. Y., Liu, X. J., & Rui, Y. (2015). Automatically solving number word problems by semantic parsing and reasoning. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 1132–1142.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A.,

Garriga-Alonso, A. & et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv Preprint*, arXiv:2206.04615.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., & Wei, J. (2023). Challenging big-bench tasks and whether chain-of-thought can solve them. In: *Proceedings of the Findings of the Association for Computational Linguistics*. Stroudsburg: ACL, 13003–13051.

Wang, Y., Liu, X. J., & Shi, S. M. (2017). Deep neural solver for math word problems. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 845–854.

Yuan, Z., Yuan, H. Y., Tan, C. Q., Wang, W., & Huang, S. F. (2023). How well do large language models perform in arithmetic tasks? *arXiv Preprint*, arXiv:2304.02015.

Zhan, B. J., Guo, T., Li, X. Y., Hou, M. L., Liang, Q. R., Gao, B. Y., Luo, W. Q., & Liu, Z. T. (2024). Knowledge tracing as language processing: A large-scale autoregressive paradigm. In: *Proceedings of the 25th International Conference on Artificial Intelligence in Education*. Berlin: Springer-Verlag, 177–191.

Zhang, X. T., Li, C. Y., Zong, Y., Ying, Z. Y., He, L., & Qiu, X. P. (2023). Evaluating the performance of large language models on GAOKAO benchmark. *arXiv Preprint*, arXiv:2305.12474.

Zhao, W., Shang, M. Y., Liu, Y., Wang, L., & Liu, J. M. (2020). Ape210K: A large-scale and template-rich dataset of math word problems. *arXiv Preprint*, arXiv:2009.11506.

Zheng, Y., Li, X. Y., Huang, Y. Y., Liang, Q. R., Guo, T., Hou, M. L., Gao, B. Y., Tian, M., Liu, Z. T., & Luo, W. Q. (2024). Automatic lesson plan generation via large language models with self-critique prompting. In: *Proceedings of the 25th International Conference on Artificial Intelligence in Education*. Berlin: Springer-Verlag, 163–178.

Zhong, W. J., Cui, R. X., Guo, Y. D., Liang, Y. B., Lu, S., Wang, Y. L., Saied, A., Chen, W. Z., & Duan, N. (2024). AGIEval: A human-centric benchmark for evaluating foundation models. In: *Proceedings of the Findings of the Association for Computational Linguistics*. Stroudsburg: ACL, 2299–2314.