2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

GSM-AIV: Exposing the Fragile Boundaries of Mathematical Reasoning in LLMs through Contextual Recomposition

Hualing Liu*[1], Zilong Zhang[1], Yingxin Hong[1], Yiwei Guo[1], Shiqin Gong[1], and Mengkai Wang[1]

[1] School of Statistics and Information, Shanghai University of International Business and Economics, 201620, China
liuhl@suibe.edu.cn
* Corresponding author

Abstract. This study reveals the semantic contextual vulnerability of large language models in solving mathematical problems. By constructing GSM-AIV, an Algebraic Isomorphic Variants dataset of GSM8K, we find that large language models such as Mistral-7b have an average accuracy improvement of 5.65 percentage points under the condition of retaining the mathematical logical chain but changing the problem context. This phenomenon implies that mathematical reasoning in LLMs relies heavily on surface semantic patterns rather than deep mathematical understanding. We further propose the template-induced bias and attention entropy reduction hypotheses to argue for the phenomenon of loose coupling between the mathematical reasoning ability of the models and the semantic scenarios, which provides a new theoretical perspective on the design of evaluation frameworks.

Keywords: Large Language Model, Natural Language Processing, Mathematical Reasoning.

1 Introduction

With the breakthrough of Large Language Models (LLMs) in natural language understanding and reasoning tasks [1,2], they have shown unprecedented potential in the field of mathematical problem solving. Recent studies have shown [3,4] that even models with less than 10 billion parameters can achieve more than 80% accuracy in authoritative mathematical benchmarks such as GSM8K after enhanced fine-tuning. This development has led researchers to focus on whether LLMs have truly mastered human-level mathematical reasoning or whether they are reaching superficial performance metrics through data-driven pattern recognition strategies [5].

Existing research on the assessment of LLM's mathematical reasoning ability is usually based on the decoupling assumption of mathematical ability, i.e., the assumption that problem solving performance should be independent of the specific problem formulation. As a result, existing benchmarking perturbation studies have focused on

direct modifications of mathematical structures (e.g., numerical substitutions or operator changes) [6], or have been limited to shallow adjustments of semantic representations (e.g., changing "per day" to "daily") [7]. This assessment paradigm is fundamentally limited: the experimental design fails to effectively decouple the mathematical logic kernel from the context of the problem, resulting in the measurement of the model's mathematical reasoning ability being consistently contaminated by superficial semantic features.

This study reveals a key cognitive divergence: when confronted with mathematically meaningful isomorphic problems, the output of LLMs exhibits significant context sensitivity. By constructing a dataset of Algebraic Isomorphic Variants, we observe significant counter-intuitive phenomenon after implementing a complete mathematical structure preserving transformation on the GSM8K benchmark: a systematic evaluation of mainstream open-source LLMs reveals that models with small parameters (less than 10 billion) exhibit "inverse robustness ", with an average accuracy improvement of 5.65 percentage points after their problem context is reshaped. This directly subverts the classical expectation that the inference performance of LLMs degrades with the imposition of dataset perturbations, implying that the mathematical reasoning ability of LLMs is deeply bound to specific semantic templates rather than being built on an abstract comprehension of the mathematical kernel.

The main contributions of this paper are as follows:

First, we establish a framework for generating the Algebraic Isomorphic Variants dataset (GSM-AIV) based on semantic reconstruction, which realizes systematic substitution of the problem background, entity referent and syntactic structure while strictly maintaining the integrity of the mathematical logic chain of the original problem;

Second, the study observes the counter-intuitive properties of LLMs in mathematical reasoning: when questions are subjected to semantic scenario substitutions while maintaining mathematical equivalence, the model's answer accuracy exhibits the opposite of the expected tendency to increase;

Finally, we propose the template-induced bias and attentional entropy reduction hypotheses to provide an explanation for the observed counter-intuitive phenomenon: the choice of inference paths of a LLM in parsing mathematical problems relies on the distribution of specific semantic templates, instead of strictly adhering to the intrinsic logical constraints of the mathematical structure.

The above findings provide new theoretical perspectives for the assessment of LLM’s mathematical reasoning ability, while pointing out the cognitive architectural deficiencies that still need to be addressed for the construction of truly robust reasoning models.

2 Related Work

2.1 Logical Reasoning Ability of LLM

Many studies have been conducted on the logical reasoning ability of LLM. Wu et al. found that LLM has a "Reversal Curse", i.e., it is unable to deal with the reverse logical relations (e.g., the reverse reasoning of "A→B") effectively [8]. Liu et al. pointed out

that in complex reasoning scenarios, such as mathematical induction and inversion, LLMs often show mismatches between assumptions and conclusions, and difficulty in maintaining logical consistency [9]. Wu et al. found that LLMs show obvious logical breaks in deductive reasoning with more than 3 steps in a benchmark test of syllogistic reasoning [10].

While the above studies discussed the inherent limitations of LLMs in performing logical reasoning tasks from different perspectives, our research focuses on the field of mathematical reasoning and reveals the logical reasoning vulnerability of LLMs by modifying the semantic scenarios.

2.2 GSM8K Benchmark

GSM8K is a dataset containing elementary school level mathematical problems for evaluating the stepwise reasoning ability of models, which has been used as a core benchmark for mathematical reasoning evaluation in many studies [11]. Among closed-source models, GPT-4 achieves 94.8% accuracy on GSM8K [12], while open-source models such as Qwen2.5-Math and MAmmoTH2 can similarly achieve competitive accuracy by augmenting their mathematical reasoning capabilities in the post-training phase [13,14].

GSM-IC introduces irrelevant contexts on top of GSM8K, tests the model resistance to interference, and found that models with higher accuracy for the original problem perform significantly lower on GSM-IC [15].GSM-Plus further extends GSM8K by introducing more complex scenarios (e.g., multi-language, symbolic interference, and multi-step reasoning) to comprehensively evaluate model robustness [7]. Experiments show that the human-model gap widens significantly on GSM-Plus. GSM-Symbolic extracts variables (e.g., entity name, numerical value) and conditions from GSM8K problems, generates parameterizable templates, and dynamically generates problems of varying difficulty, and similarly observes pattern-matching for which the model is highly dependent on the pre-training data [6].

Unlike the above work, our research focuses on modifying the semantic context rather than the mathematical logic chain itself, revealing the impact of the semantic scenario on the LLM's logical reasoning ability.

3 The GSM-AIV Dataset

In order to construct algebraic isomorphic variants dataset, we design a generative framework based on the logical chaining of problem solving formulations. The core idea of the framework is to generate mathematical problems that are logically equivalent in terms of solution but different in semantic description by keeping the mathematical logic structure unchanged and replacing only the problem background description:

$$Q_{original} = (V, E, \varphi) \rightarrow Q_{AIV} = (V', E, \varphi) \quad (1)$$

where V denotes the semantic entities, E denotes the arithmetic relations, and φ denotes the numerical value corresponding to V.

3.1 Data Acquisition and Preprocessing

First, 1000 pairs of problem-solving processes (Q, S_q) are randomly sampled from the training and test sets of the GSM8K dataset at a ratio of 8:2 and two-stage filtering is performed:

Equation exclusion. Since problems containing equations are more difficult to be understood by the generative model, we expect problems to be generated by complete mathematical logic chains that do not contain unknown variables. Specifically, we remove all samples that satisfy $\exists t \in S_q, t \in E_{equations}$ by a pattern matching algorithm, where $E_{equations}$ is the set of expressions that contain an unknown variable (typically *x*).

Mathematical logic chain extraction. Structured parsing of retained samples containing only purely arithmetic paths to construct formal representations:

$$C_q =< \{e_i\}_{i=1}^{n}, r > \tag{2}$$

Where e_i denotes the arithmetic expression at step i and *r* denotes the final numerical solution. Specifically, since the original GSM8K dataset is labeled with arithmetic expressions and final numerical solutions in special forms ("<<>>" and "####") in S_q, we can extract these elements from the natural language form of S_q by regular expressions, and then compare the result of the last step of the arithmetic expression e_n with the final numerical solution r. We consider the C_q that both are equivalent as the "correct" problems and keep it.

3.2 Prompt Engineering Module

After obtaining the complete and correct mathematical logic chain, it is necessary to embed the chain C_q into a predefined context-constrained prompt template, and then the large language model generates new math problems with exactly the same solution logic but different semantic contexts based on the prompt. Specifically, we called the DeepSeek-R1-671B model through the API with the hyperparameters of temperature=0.6 and top-p=0.6, which can better balance the generation diversity and semantic controllability at this situation, and meet the task requirements. The prompt template is shown in Fig. 1.

Given arithmetic operations and result:
Steps: {e_1}; {e_2}; ... ; {e_n}
Result: {r}
Generate a math problem with a backstory that:
1. Requires EXACTLY these operations in order;
2. Matches elementary school difficulty.
DO NOT show the solution process and final answer in the question. Output format:
Problem: [Your Question]

Fig. 1. Prompt template of the problem generation

3.3 Post-processing Validation Module

In order to ensure the quality of the GSM-AIV dataset, further validation and filtering processing of the data generated by LLM is required. Specifically, we have established a three-level progressive validation method:

Contextual Dissimilarity Assessment. In order to assess whether the new problems generated by LLM are semantically and structurally different from the original problems, we define the contextual dissimilarity score (CDS):

$$CDS = 1 - \frac{ROUGE(q,q') + BERTScore(q,q')}{2} \tag{3}$$

In this formula, The ROUGE score is responsible for capturing the surface structure match between the new problem and the original problem text, and the BERT score is responsible for measuring the semantic equivalence between the new problem and the original problem text. By combining the two metrics, CDS is able to synthesize the surface textual similarity and deep semantic alignment to filter out new problems that are sufficiently different from the original problem.

We require all the new problems generated by LLM to satisfy CDS>0.75, and the distribution of ROUGE scores and BERT scores of the problems after filtering is shown in Fig. 2. It can be seen that the ROUGE scores of the filtered algebraic isomorphic problems are distributed in the range of 0 to 0.25, and the BERT score is distributed in the range of 0.225 to 0.450, which are kept at a low level, indicating that the new problems generated by the LLM are sufficiently different from the original topics in terms of semantic context. The next step of validation is performed on these problems.

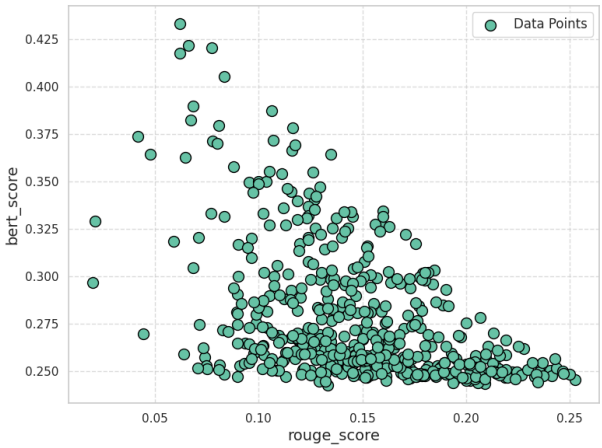

Fig. 2. The distribution of ROUGE scores and BERT scores of the problems

Cognitive Complexity Alignment. To eliminate cognitive bias due to semantic reconstruction, we constructed a GPT-4-based complexity assessment agent that uses standardized prompt to ensure that the mathematical complexity of the new problem matches that of the original problem. The template of the prompt for the complexity assessment agent is shown in Fig. 3.

Compare problem A and B on mathematical complexity:
A: [Original Question]
B: [Generated Question]
Output '1' if B is strictly harder, '-1' if A is strictly harder, '0' if A and B are of equal difficulty

Fig. 3. Prompt template for the complexity assessment agent

Problems where the agent is judged to be '0', i.e., the original problem is the same in mathematical complexity as the new problem, are retained for the next step of validation.

Logical Equivalence Validation. we recruited three graduate students enrolled in math-related majors as experts to independently perform manual validation of the new problems generated by LLM. We provided each expert with the mathematical logic chain, the original problem in the GSM8K dataset corresponding to the mathematical

logic chain, and the new problem generated by LLM based on the mathematical logic chain. After that, We asked the experts to judge whether the solution logic of the new problem conforms to the given mathematical logic chain, and whether the semantic context of the new problem is different from that of the original problem. The new problem is retained if and only if all three experts agree that the new problem and the original problem are algebraic isomorphic problems with different semantic scenarios.

After the above three steps of validation, a total of 789 valid algebraic isomorphic variant problems were obtained, and the set was used as the GSM-AIV dataset. The original 789 questions with the same mathematical logic chain in the GSM8K dataset were used as a control set. The Table 1. shows an example of GSM-AIV. The novelty of GSM-AIV is that its solution logic and final answers are identical to the corresponding questions in GSM8K, but the semantic context of each question is rewritten. By testing GSM-AIV against GSM8K, we can explore the semantic coupling of large language models in the mathematical reasoning process.

Table 1. An example of GSM-AIV

Dataset	Example
GSM-AIV	**Four friends** decided to gather apples from their backyard trees to donate. Each friend picked **6 apples**. They combined their harvest and planned to distribute the apples equally among **3 local food banks**. How many apples did each food bank receive?
GSM8K	It takes Jerome **6 hours** to run the trail around the park and it takes Nero **3 hours**. If Jerome runs at **4 MPH**, what speed (in MPH) does Nero run in the park?
Mathematical Logic Chain	e_1:4*6=24 e_2:24/3=8 r:8

4 Experiments

4.1 Experimental Setup

We systematically reviewed 15 mainstream models on the GSM-AIV dataset, covering a variety of parameter sizes ranging from billions to hundreds of billions. Specifically, we consider closed-source foundation models, i.e., GPT-4o and GPT-3,5-Turbo, open-source foundation models, i.e., Mistral, Llama-3 [16], Gemma [17,18], and Phi3 [19], as well as open-source enhanced fine-tuned models specifically designed for mathematical reasoning, i.e., Qwen-2.5-Math [3] and MAmmoTH-Plus [4]. For closed-source models, we experiment by calling APIs; for open-source models, we experiment after deploying them locally. With using the 0-shot CoT prompt method [20], each model completes five rounds of independent inference processes on both datasets and the accuracy is averaged. All models' hyperparameters are taken as the recommended optimal hyperparameters (typically the temperature is set to 0.7~1 and top-k is set to 50~100) to ensure that the model achieves its own optimal results.

4.2 Experimental Results

Table 2 shows the results of the comparative evaluation of the subjected models, where most of the models show significant performance oscillations after semantic refactoring, and small models (less than 10 billion parameters) exhibit particular performance gain phenomenon. Our main findings are summarized below.

Table 2. The accuracy of subjected LLMs on GSM8K and GSM-AIV.

Model name	Number of parameters	GSM-AIV Acc.(%)	GSM8K Acc.(%)	ΔAcc.(%)
GPT-3.5-turbo	unknown	86.06	85.93	0.13
GPT-4o-0513	unknown	86.44	87.71	-1.27
Deepseek-v3	671b	92.90	94.80	-1.9
Llama3	70b	86.69	89.99	-3.3
Llama3	8b	73.26	67.30	5.96
Qwen2	7b	79.09	79.72	-0.63
Mistral-v0.1	7b	52.47	37.90	14.57
Mistral-v0.3	7b	58.30	42.21	16.09
Gemma2	9b	82.38	79.97	2.41
Gemma3	12b	84.41	84.66	-0.25
Gemma3	4b	78.58	72.37	6.21
Phi3	14b	69.58	63.88	5.7
Phi3	3.8b	22.81	21.80	1.01
Qwen-2.5-math	7b	95.69	94.0	1.69
Mammoth-plus	8b	75.16	71.61	3.55

Anomalous Gain Patterns in Smaller LLMs. LLMs with less than 10 billion parameters are significantly more accurate on GSM-AIV than on the original GSM8K. the Mistral-7B family of models performs particularly evident, with version 0.1 improving accuracy on GSM-AIV by up to 14.57 absolute percentage points, and version 0.3 by up to 16.09 absolute percentage points. The phenomenon is generalizable to smaller LLMs - 89% (8/9) of the models with less than 10 billion parameters satisfy $\Delta Acc>0$, with an average gain of 5.65%, which is significantly negatively correlated with the number of parameters in the model.

Constraint on Counter-intuitive Gain. The gain phenomenon gradually dissipates when the number of parameters exceeds a critical threshold (~20B) (e.g., ΔAcc of -3.30 for Llama-3.3 model with 70 billion parameters). The non-significant fluctuation of Acc in the closed-source model group (GPT-4o and GPT-3.5-Turbo) also validates the pattern. This phenomenon implies that the performance gain may originate from the overfitting of the smaller LLMs to the potential semantic template, and that the model with a larger number of parameters is better able to abstract mathematical logical chains from different semantic contexts. Note that the magnitude of gain of the enhanced fine-tuned model (e.g., Qwen-2.5-Math) is significantly lower ($\Delta Acc = 1.69$) than that of the foundation models of the same parameter size due to the optimization of the specialized training data.

4.3 Analysis

In order to explain the phenomenon of anomalous gain of smaller LLMs found in 4.2, we propose the template-induced bias and the attentional entropy reduction hypotheses to try to explain the reasons behind this phenomenon.

Template-Induced Bias. LLMs may be strongly guided by the input of mathematical logic chains when generating algebraic isomorphic variant problems, and involuntarily reproduce patterns with significant surface feature associations between inputs and outputs. We performed pattern reproducibility detection by using the self-BLEU metric:

$$Self - BLEU = \frac{1}{N}\sum_{i=1}^{N} BLEU\left(q_i, \{q_j\}_{j\neq i}\right) \tag{4}$$

The Self-BLEU score of GSM-AIV was 0.458 and that of the original GSM8K was 0.284, as well as the p-value of the Mann-Whitney U-test was $1.34\text{^}10^{-127}$, suggesting that the difference in Self-BLEU scores between the two datasets was significant.

The fact that GSM-AIV had a high Self-BLEU score suggests that the generated problems were more vocabulary and syntactically consistent, suggesting that the model follows a narrow set of templates to generate new problems. The structural monotonicity of AIV was also detected by human experts, for example, words such as “school”, “Lila”, “apples”, and “cookies” appeared significantly more frequently than in the original GSM8K dataset, suggesting that these words may have a template-induced bias and are more likely to be outputted in contexts that are strongly associated with mathematical logic.

Attention Entropy Reduction. We assume that when the amount of information describing a natural language problem is reduced, the model's inference resources are allocated more efficiently. We use Measure of Textual Lexical Diversity (MTLD) as an information complexity proxy:

$$MTLD = \frac{N}{\sum_{i=1}^{k}\frac{L_i}{F}} \tag{5}$$

Where N is the total number of words in the text, k is the total number of sequences that satisfy Type-Token Ratio (TTR) ≥ 0.72, L_i is the length of the i-th sequence, and F is the factor size (here taken as 10).

To further improve the stability of the results, the MTLD values were calculated separately for the text from forward and backward directions and then the average of the two was taken:

$$MTLD_{final} = \frac{MTLD_{forward}+MTLD_{backward}}{2} \tag{6}$$

The MTLD score for the GSM-AIV was 1.639 and the MTLD score for the original GSM8K was 1.683, as well as the p-value of the Mann-Whitney U test was $3.02\text{^}10^{-4}$, suggesting that the difference in MTLD scores between the two datasets was significant.

Lower MTLD score indicates that GSM-AIV has more keyword repetition and less descriptive words, which may lead to the subjected LLM capturing the core of mathematical logic more easily: the model, due to the reduction of input language complexity, allocates its attention to focus more on parsing the mathematical logic (e.g., enhanced attention to numerical values and operators), which improves reasoning accuracy. Through word frequency detection, we found that words such as “each”, “total”, “equally”, and “remaining” appeared significantly more frequently than in the original GSM8K data dataset, and the occurrence of these words may induce the model to allocate more attention to mathematical logic.

In summary, the synergistic effect of the template-induced bias of the GSM-AIV problems, which promotes deterministic inference paths, and its low linguistic complexity, which enables optimal allocation of attention, leads to a counter-intuitive increase in the accuracy of LLMs with a smaller number of parameters on GSM-AIV. This may suggest that the "mathematical reasoning ability" of existing LLMs, especially smaller models, are essentially computational short-circuit connectivity driven by the semantic-computational template cooccurrence pattern in the pre-training data, rather than causal reasoning based on algebraic structures.

5 Conclusion

By constructing the algebraic isomorphic variant dataset GSM-AIV on the GSM8K dataset, this study finds that LLMs exhibit a unique semantic path dependency in mathematical reasoning: smaller scale models generally exhibit counter-intuitive performance gains when the problem is kept mathematically isomorphic but reconstructed semantic scenarios. This counter-intuitive phenomenon may stems from a shallow binding mechanism between computational processes and semantic representations - algebraic equivalence problems generated via structured prompt exhibit a significant template-induced bias, and their lower lexical diversity creates an attentional entropy reduction effect, forcing the models to prioritize parsing high-frequency arithmetic keywords.

This study provides a new theoretical framework for the evaluation of mathematical reasoning ability: current traditional mathematical benchmarks are at risk of representational bias, and truly robust algebraic reasoning models need to achieve deep decoupling of semantic representations and mathematical logic. However, this study still has some limitations: The construction of the algebraic isomorphic variant dataset is limited to GSM8K dataset, which only explores the simpler mathematical issues. As well, we have not explored the internal mechanisms of the model enough. Future work will extend the algebraic isomorphic variant dataset construction to more domains and more fine-tuned expert models to confirm whether this is a widespread phenomenon and explore its underlying causes, as well as an algebraic invariant-driven training paradigm to break through the current mathematical logical reasoning bottleneck that relies on generalizable semantic templates.

References

1. Collins K M, Jiang A Q, Frieder S, et al. Evaluating language models for mathematics through interactions[J]. Proceedings of the National Academy of Sciences, 2024, 121(24): e2318124121.
2. Yu Z, He L, Wu Z, et al. Towards better chain-of-thought prompting strategies: A survey[J]. arXiv preprint arXiv:2310.04959, 2023.
3. Shao Z, Wang P, Zhu Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[J]. arXiv preprint arXiv:2402.03300, 2024.
4. Wang P, Li L, Shao Z, et al. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 9426-9439.
5. Davoodi A G, Davoudi S P M, Pezeshkpour P. Llms are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of llms[J]. arXiv preprint arXiv:2406.05194, 2024.
6. Mirzadeh I, Alizadeh K, Shahrokhi H, et al. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models[J]. arXiv preprint arXiv:2410.05229, 2024.
7. Li Q, Cui L, Zhao X, et al. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers[J]. arXiv preprint arXiv:2402.19255, 2024.
8. Wu D, Yang J, Wang K. Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models[J]. Patterns, 2024, 5(9).
9. Liu H, Ning R, Teng Z, et al. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4[J]. CoRR, 2023.
10. Wu Y, Han M, Zhu Y, et al. Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 2347-2367.
11. Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems[J]. arXiv preprint arXiv:2110.14168, 2021.
12. Tan W, Chen D, Xue J, et al. Teaching-Inspired Integrated Prompting Framework: A Novel Approach for Enhancing Reasoning in Large Language Models[J]. arXiv preprint arXiv:2410.08068, 2024.
13. Zhang Z, Zheng C, Wu Y, et al. The lessons of developing process reward models in mathematical reasoning[J]. arXiv preprint arXiv:2501.07301, 2025.
14. Yue X, Zheng T, Zhang G, et al. Mammoth2: Scaling instructions from the web[J]. Advances in Neural Information Processing Systems, 2024, 37: 90629-90660.
15. Shi F, Chen X, Misra K, et al. Large language models can be easily distracted by irrelevant context[C]//International Conference on Machine Learning. PMLR, 2023: 31210-31227.
16. Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.
17. Team G, Riviere M, Pathak S, et al. Gemma 2: Improving open language models at a practical size[J]. arXiv preprint arXiv:2408.00118, 2024.
18. Team G, Kamath A, Ferret J, et al. Gemma 3 technical report[J]. arXiv preprint arXiv:2503.19786, 2025.
19. Abdin M, Aneja J, Awadalla H, et al. Phi-3 technical report: A highly capable language model locally on your phone[J]. arXiv preprint arXiv:2404.14219, 2024.
20. Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[J]. Advances in neural information processing systems, 2022, 35: 22199-22213.