Atomic Thinking of LLMs: Decoupling and Exploring Mathematical Reasoning Abilities

Jiayi Kuang[1], **Haojing Huang**[2], **Yinghui Li**[2,†,*], **Xinnian Liang**[3], **Zhikun Xu**[4]
Yangning Li[2], **Xiaoyu Tan**[5], **Chao Qu**[6], **Meishan Zhang**[7], **Ying Shen**[1,8,†], **Philip S. Yu**[9]
[1]Sun Yat-sen University, [2]Tsinghua University, [3]ByteDance Inc.
[4]Arizona State University, [5] Tencent Youtu Lab, [6]Fudan University
[7]Harbin Institute of Technology (Shenzhen), [8]Pengcheng Laboratory
[9]University of Illinois Chicago

Abstract

Large Language Models (LLMs) have demonstrated outstanding performance in mathematical reasoning capabilities. However, we argue that current large-scale reasoning models primarily rely on scaling up training datasets with diverse mathematical problems and long thinking chains, which raises questions about whether LLMs genuinely acquire mathematical concepts and reasoning principles or merely remember the training data. In contrast, humans tend to break down complex problems into multiple fundamental atomic capabilities. Inspired by this, we propose a new paradigm for evaluating mathematical atomic capabilities. Our work categorizes atomic abilities into two dimensions: (1) **field-specific abilities across four major mathematical fields**, algebra, geometry, analysis, and topology, and (2) **logical abilities at different levels**, including conceptual understanding, forward multi-step reasoning with formal math language, and counterexample-driven backward reasoning. We propose corresponding training and evaluation datasets for each atomic capability unit, and conduct extensive experiments about how different atomic capabilities influence others, to explore the strategies to elicit the required specific atomic capability. Evaluation and experimental results on advanced models show many interesting discoveries and inspirations about the different performances of models on various atomic capabilities and the interactions between atomic capabilities. Our findings highlight the importance of decoupling mathematical intelligence into atomic components, providing new insights into model cognition and guiding the development of training strategies toward a more efficient, transferable, and cognitively grounded paradigm of "*atomic thinking*".

1 Introduction

In recent years, as Large Language Models (LLMs) have achieved remarkable performance in language understanding [1–7], visual perception [8–15], complex reasoning [16–21] agentic intelligence [22–25] and honesty [26–29], mathematical reasoning has emerged as a key focus for LLM cognitive abilities [30–33]. Mathematics, as a fundamental reasoning task, offers both verifiable answers and a wide range of difficulty levels, drawing increasing research attention [34–36]. Recent studies, particularly those involving reasoning models such as OpenAI's o1 and DeepSeek-R1, have demonstrated strong mathematical performance, achieving impressive results on many challenging benchmarks [37–39].

[*]Yinghui Li is the project leader.
[†]Correspond to Yinghui Li (liyinghuihhh@gmail.com), Ying Shen (sheny76@mail.sysu.edu.cn). Ying Shen is the corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

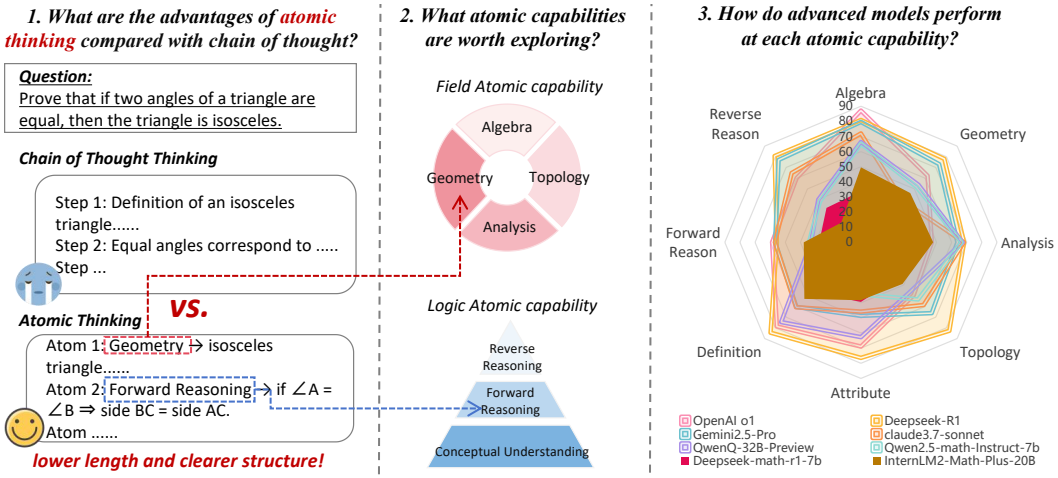

Figure 1: This figure illustrates an overview of our atomic thinking. It compares thought chains and atomic thinking, highlighting the efficiency of atomic thinking. Next, it shows the atomic capabilities we focus on. Finally, it provides the performance of advanced models in every atomic capability.

Current approaches enhance mathematical performance by scaling up training data, incorporating diverse mathematical problems, and complex reasoning paradigms [40, 41, 31]. WizardMath [32] synthesizes complex multi-field data via varied math instructions, while models like OpenAI o1 train extended chains of thought to perform higher-level reasoning. As the complexity of mathematical problems increases, some studies, such as Lean4 [42] and Lean-STaR [43], leverage formal mathematical languages to mitigate ambiguity in natural language multi-step reasoning, as well as enhancing the reward model for problem-solving process [44–46]. However, as model size and data scale approach saturation, these training paradigms are facing challenges:

- *Do models truly grasp mathematical concepts and inference patterns, or are they merely memory problems by chain of thought reasoning training?*
- *Is there a more fundamental cognitive atom that can break through the current paradigm, and* ***what are the advantages of this atomic thinking compared with the chain of thought?***

Figure 1 compares the differences between the two paradigms. Existing math reasoning strategies, such as chain-of-thought and tree-of-thought, tend to rely heavily on sequential context [34, 47–49]. This not only leads to inefficient use of computational resources but also introduces noise through excessive self-correction in long reasoning chains. In contrast, human reasoning typically decomposes complex problems into atomic problems, solving them incrementally and integrating only the essential information for subsequent steps, which is referred to as *atomic thinking* [50]. This atomic thinking paradigm promotes flexible and structural reasoning, enabling more efficient problem-solving like data probe [51–53]. Thus, decoupling the mathematical atomic abilities of LLMs is not only essential for assessing their cognitive depth but also key to transitioning from the current "question drilling" paradigm to a *"atomic thinking"* framework. So we wonder ***what atomic capabilities are worth exploring?***

Current mathematical benchmarks mostly assess models' accuracy in end-to-end problem solving [54–56], offering little insight into a systematic assessment of atomic capabilities. To address this, we propose a novel framework for exploring mathematical atomic abilities, encompassing both field and logical reasoning capabilities, with an emphasis on ensuring minimal overlap between different atomic units while maintaining broad coverage of mathematical tasks. For field atomic abilities, we draw inspiration from modern mathematics and construct four foundational fields: **algebra, geometry, analysis, and topology**. For logical reasoning abilities, we reference cognitive psychology in math reasoning to define three core capabilities: (1) **Conceptual Understanding**, which grasps math definitions and axioms; (2) **Forward Reasoning with Formal Math Language**, which conducts rigorous, multi-step reasoning using symbolic systems; (3) **Counterexample-driven Backward Reasoning**, requiring constructing counter-examples and leveraging backward reasoning. For each atomic ability, we construct training and testing data, ensuring interpretability and isolating cross-

ability. We conduct the evaluation experiments as shown in Figure 1 to explore: ***how do advanced models perform at each atomic capability?***

Beyond decoupled atomic capability evaluation, we also conduct composite experiments across reasoning levels and varied atomic abilities to investigate how atomic abilities influence each other. Our findings on all the results reveal several key insights:

- **Field-level performance**: LLMs perform better in algebra and analysis, while struggling in geometry and topology. Interestingly, models exhibit atypical behavior in topology, performing worse on easier tasks yet better on harder ones.
- **Logical reasoning abilities**: Larger LLMs exhibit stronger conceptual understanding, likely due to superior pretraining memory. However, even advanced commercial models struggle with constructing counterexamples, indicating a gap in backward reasoning skills.
- **Cross-field interaction**: Training on low-difficulty data can hinder high-level skill expression in some fields. Notably, activating algebraic abilities significantly improves performance in other fields, which is often more than direct training in the target field.
- **Cross-Logic interaction**: Conceptual understanding enhances other reasoning abilities and field atomic abilities. Surprisingly, training solely on definition completion tasks suffices to stimulate high-level ability, outperforming models trained on more complex data. This reveals the supporting value of conceptual understanding in mathematical training.

2 Related work

Mathematics-enhanced large language models Mathematical reasoning has become a key focus in exploring the upper bounds of LLMs' cognitive capabilities [57, 58]. Unlike general-purpose models such as GPT-4 [37] and Gemini [30], math-enhanced LLMs emphasize field-specific strategies, including data augmentation, pretraining, fine-tuning [59], and reinforcement learning on large-scale mathematical corpora [36]. **WizardMath** [32] synthesizes diverse math data through instruction generation and leverages RLHF and process supervision. **NuminaMath** [60] adopts Tool-Integrated Reasoning (TIR) to generate math data, including questions with fine-grained solutions. **Qwen2.5-Math** [31] fine-tunes Qwen2.5 on proprietary high-quality math data. **InternLM2-Math** [35] enhances logical rigor by integrating formal mathematical language, code interpreters, and theorem proving in Lean4. **Deepseek-Math-rl** [34] focuses on data engineering and efficient RL training.

Mathematical benchmarking Due to their objective correctness and structured difficulty [61, 62], math tasks are ideal for evaluating LLMs [54, 63, 64, 55, 56]. Benchmarks like **MATH** [65] and **GSM8K** [66] test high school and elementary-level reasoning and have become standard. To meet the demands of advanced models, more challenging datasets such as **OlympiadBench** [67] target competition-level exams. Formal theorem proving benchmarks like [68] **Putnam Bench**, **CoqGym** [63], and **MiniF2F** [64] further assess logical reasoning with tools like Coq and Lean. However, most benchmarks assess end-to-end question-solving performance without decomposing tasks into **atomic reasoning abilities**. Our work aims to decouple and analyze these atomic abilities and their interactions to support finer-grained reasoning evaluation and lightweight but effective training.

3 Atomic capability decoupling and interaction

Inspired by the atomization of human cognition, we propose decoupling the mathematical capabilities of LLMs into atomic abilities. We categorize atomic capabilities into two major types, ensuring that each capability is disentangled from the others while jointly covering a wide range of mathematical tasks. We construct corresponding training and test sets to evaluate the model performance of different atomic abilities. Beyond evaluating individual capabilities, we further investigate their interactions. Specifically, we explore how stimulating one atomic capability may affect others, offering insights into how such interactions can be leveraged to enhance targeted atomic abilities and promote compositional problem-solving strategies.

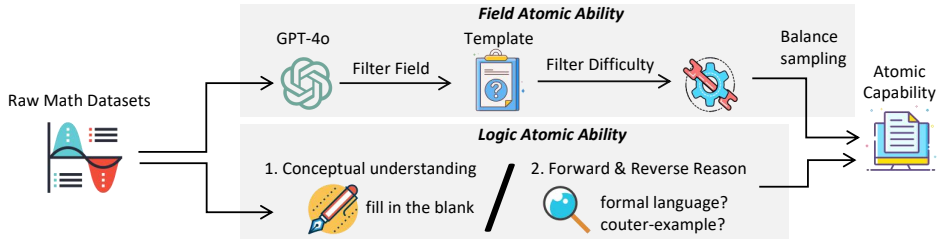

Figure 2: This figure illustrates our data construction procedure.

3.1 Atomic capability design

Field atomic capabilities We refer to the core field division of modern mathematics. It is worth noting that if we divide the capabilities too finely (e.g., dividing them into more than ten types), the atomic units will be too loose and the interaction between atomic abilities will be too complex, which may not be a particularly significant correlation. Therefore, we finally divide the field's atomic abilities into **Algebra**, **Geometry**, **Analysis**, and **Topology**. To analyze the influence of difficulty levels in a more fine-grained aspect, we further divide each field into low difficulty (level 1) and high difficulty (level 2). Similarly, we divide it into two difficulty levels to avoid complex interactions.

Logical reasoning atomic capabilities For logic reasoning atomic ability, we refer to the human cognitive mode, deconstruct the general process of complex reasoning, and identify core abilities:

1. **Conceptual understanding** is a fundamental, which includes *definition identification* and *attribute description*. The definition identification task requires the LLM to complete the name of the corresponding definition in a statement. The property description task requires completing the detailed description in a mathematical definition, including the premises, the conditions, the key words, and parameters.
2. **Forward reasoning with formal language**: With reference to human cognitive processes, we designate forward multi-step reasoning as a higher-level logical atomic ability. Since natural language reasoning is often confronted with vague proofs and uncritical assumptions, we emphasize the forward reasoning using formal mathematical language.
3. **Backward reasoning with counterexamples**: In addition to step-by-step forward multi-step reasoning, backward reasoning with counterexamples is also a very important ability in mathematical reasoning. By skillfully constructing appropriate counterexamples, proofs can be effectively accomplished that are difficult to perform directly with forward reasoning.

It is worth noting that we do not consider computational abilities. This is because we are mainly concerned with the logical reasoning ability of mathematical reasoning. Computational capability is difficult to decouple from the reasoning process. In addition, when computations are needed, LLM invoking relevant mature computational tools would be more efficient and accurate.

3.2 Data construction for capability evaluation

To decouple each atomic capability, we construct the training and testing data, which contributes to evaluation and further exploration of the interaction. The overview of our data construction is shown in Figure 2. More data statistics and examples can be found in Appendix A.

For field capabilities, we collect data from current benchmarks such as MATH [65], GSM8K [66], Gaokao-Bench [69], OlympiadBench [67], AIME, MMLU [70], and DeepMath [71]. Using a combination of template matching and LLM-assisted annotation, we reclassify problems into four fields. When there are original field labels in the raw data, LLM takes more account of the primitive labels to better align with human thinking. Difficulty annotation is done by template matching. We take into account the data sources of the original questions (e.g., primary and secondary school questions or Olympiad questions), the original difficulty labeling information, and we emphasize aligning the difficulty classifications of different fields. After that, we randomly sample the data to ensure that the number of each field is relatively balanced. Finally, we randomly divide the training set and test set with a ratio of 3:1. For **conceptual understanding**, we extract math definitions and axioms from NaturalProofs [72] and generate fill-in-the-blank questions. For **forward reasoning**, we

collect questions and proofs with formal language such as LeanWorkbook, and filter the questions with definite answers. For **backward reasoning**, we use a counterexample-driven reasoning statement from CounterMath [73]. Given data scarcity, we maintain a 1:1 train-test split for logic atomic capability to ensure evaluation robustness.

3.3 Atomic capability interaction

In addition to decoupling individual atomic capabilities, we also try to explore the correlation and interaction between atomic capabilities. One of the most basic strategies is trying to stimulate a certain atomic capability and observe whether it has an impact on other atomic capabilities. This will provide insight for subsequent research on how to specifically stimulate the required atomic capabilities. Therefore, we explore the following atomic capability interactions:

- **Cross-difficulty Interaction:** We investigate how training on Level 1 or Level 2 tasks within a field affects performance at the other level.
- **Cross-field Interaction:** We train on tasks from one field and evaluate transfer effects to others, particularly among related or complementary fields.
- **Logical Capability Interactions:** We examine whether conceptual understanding can support high-level ability, improving forward or backward reasoning, and whether forward and backward reasoning mutually reinforce each other. Additionally, we assess whether high-level reasoning enhances foundational understanding.
- **Reasoning-to-field Interaction:** Given the abstract nature of some fields (e.g., topology and analysis), we test whether improving logical reasoning capabilities can enhance performance in field-specific tasks, especially those requiring conceptual abstraction.

4 Experimental settings

4.1 Decoupled atomic capability evaluation

Baselines We evaluate a diverse set of LLMs with a focus on mathematical reasoning. Open-source models include **Deepseek-Math-7B-RL** [34], **Eurus-2-7B-PRIME** [74], **Qwen2.5-Math-7B-Instruct** [31], **NuminaMath-7B-TIR** [60], **InternLM2-Plus-7B/20B** [35], **Abel-7B/13B** [36], **WizardMath-7B** [32], **Mathstral-7B**[3], **MetaMath-Mistral-7B** [40], **Xwin-Math-7B/13B** [41], **QwQ-32B**[4]. For *proprietary models*, we use **GPT-4o**, **OpenAI o1**,[5], **Deepseek-R1**[6], **Claude3.7-sonnet**[7], and **Gemini2.5-pro**[8]. This selection spans varied training paradigms, data sources, architectures, and origins (academic vs industrial). Open-source evaluations are conducted on 4× L20 48GB GPUs, while proprietary models are accessed via official APIs.

Prompts and metrics We adopt default *Chain-of-Thought* (CoT) prompting, instructing models to enclose answers in `\boxed{}` for extraction. Accuracy is computed via exact match with reference answers. For **Conceptual Understanding**, models complete missing definitions. **Forward Reasoning** requires multi-step derivations using formal mathematical language. Accuracy is used for both. For **Backward Reasoning**, we refer to previous work CounterMATH [73]: (1) F-1 score on the statement judgement, and (2) Example, Strict, Loose metric to evaluate whether the model constructs valid counterexamples. Detailed description of our prompts and metrics can be found in Appendix B.

4.2 Training for atomic capability interaction

To examine interactions among atomic abilities, we fine-tune `Qwen2.5-Math-Instruct-7B` using supervised LoRA training [75] on 4×L20 48GB GPUs, with a learning rate of 1.0e-5. Training about

[3] `https://mistral.ai/news/mathstral/`
[4] `https://qwenlm.github.io/blog/qwq-32b-preview/`
[5] `https://cdn.openai.com/o1-system-card-20241205.pdf`
[6] `https://api-docs.deepseek.com/news/news250120`
[7] `https://www.anthropic.com/claude/sonnet`
[8] `https://deepmind.google/technologies/gemini/pro/`

Table 1: Model performance on field atomic capabilities. We **bold** the optimal and underline the suboptimal of models. The low and high difficulty levels correspond to level 1 and level 2, respectively.

	Field	Algebra		Geometry		Analysis		Topology	
	Difficulty Level	Low	High	Low	High	Low	High	Low	High
	Open-source models								
Model≤7B	InternLM2-math-plus-7b	49.2	35.9	33.0	31.4	41.9	41.5	27.2	37.0
	Deepseek-math-rl-7b	52.0	33.8	35.5	39.3	44.6	37.6	22.6	33.7
	Eurus-2-7B-PRIME	50.7	33.9	36.5	32.3	45.4	32.0	36.5	23.3
	NuminaMath-7B-TIR	52.4	28.2	39.9	32.1	46.8	33.6	26.4	25.1
	MetaMath-Mistral-7B	59.6	41.4	42.4	37.7	45.2	37.9	20.6	25.8
	Mathstral-7B-v1.0	42.3	33.3	36.4	28.1	32.9	29.8	24.1	37.1
	Abel-7B-002	46.7	32.9	36.1	35.2	32.9	26.3	25.6	31.5
	Xwin-Math-7B	52.6	37.3	38.3	35.1	45.3	32.9	19.7	22.5
	Qwen2.5-math-Instruct-7b	**80.5**	**65.2**	**57.3**	**51.9**	**67.7**	**66.5**	**52.1**	**53.4**
Model > 7B	Abel-13B	63.4	47.3	42.7	41.8	40.4	35.6	30.2	24.6
	Xwin-Math-13B	78.3	51.6	49.5	46.7	55.7	54.8	46.7	38.3
	IntwenLM2-Math-Plus-20B	67.8	49.6	49.3	46.1	54.3	47.8	45.4	38.9
	QwQ-32B-Preview	**85.1**	**66.3**	**59.6**	**53.8**	**72.3**	**67.2**	**54.9**	**46.1**
	Commercial models								
	OpenAI o1	**93.6**	**87.1**	66.3	62.3	59.3	46.9	52.9	49.4
	GPT-4o	69.3	50.5	48.5	36.3	55.1	53.2	53.0	49.4
	Deepseek-r1	83.6	80.9	**76.8**	**78.0**	70.3	**68.2**	**77.7**	**82.7**
	Claude3.7-sonnet	83.5	72.0	58.0	39.0	**80.3**	68.1	56.0	58.6
	Gemini2.5-pro	80.5	79.6	70.2	73.1	67.5	63.8	68.1	66.2

different interactions adheres to a unified hardware setup and consistent hyperparameters. After training, evaluation follows Section 4.1 to ensure comparability across settings.

5 Analysis and discussion

5.1 Experimental analysis of field atomic capabilities

We evaluate several advanced models by assessing their performance across decoupled atomic abilities in distinct mathematical fields. The detailed results are presented in Table 1. We observe that:

Larger models exhibit stronger atomic capabilities Model performance varies significantly across scales. In general, larger models perform better, benefiting from greater capacity and more extensive training data. Among 7B-scale models, **Qwen2.5-math-Instruct** achieves notably superior results across all evaluated fields, even outperforming some larger models such as **InternLM2-Math-Plus-20B**. Analysis of its outputs shows that it generates longer reasoning chains, facilitating deeper logical inference and enhancing its problem-solving capabilities.

Algebra and analysis perform better Models tend to exhibit stronger mathematical atomic abilities in Algebra and Analysis, while performance in Geometry and especially Topology remains weaker. Since we analyze the training data of the open-sourced models we have evaluated, the results reveal that Geometry and Topology are significantly underrepresented in the training data. Moreover, although geometric problems are textually presented, they often require spatial or visual reasoning—an area where LLMs typically struggle. Consequently, current models demonstrate limited atomic capabilities in these fields. Improving performance in underrepresented areas, particularly Topology and Geometry, is a pressing research challenge. For example, the low performance on topology-related tasks may reflect a lack of understanding of abstract mathematical structures. A promising direction is to integrate core mathematical concepts during training to stimulate relevant atomic skills. We explore such cross-ability interactions in Section 5.3.

Table 2: Model performance on logic atomic capabilities, where Attr. and Def. are short names of the Attribute description and definition task. We **bold** the optimal and underline the suboptimal results.

		Concept		Forward Rea.	Backward Rea.			
		Attr. (Acc.)	Def. (Acc.)	Acc.	F-1	Example(%)	Strict(%)	Loose(%)
Open-source models								
Model≤7B	InternLM2-math-plus-7b	**43.2**	46.2	23.7	33.9	36.6	9.0	9.5
	Deepseek-math-rl-7b	39.4	46.5	27.6	32.2	65.9	18.9	20.6
	Eurus-2-7B-PRIME	23.1	27.9	**41.4**	37.5	64.8	28.5	32.0
	NuminaMath-7B-TIR	22.8	27.3	30.2	30.4	54.1	13.0	13.7
	MetaMath-Mistral-7B	19.6	25.6	28.6	31.0	26.5	0.4	0.7
	Mathstral-7B-v1.0	21.7	29.8	32.9	28.2	38.9	7.5	7.9
	Abel-7B-002	20.9	31.0	33.7	34.4	66.1	16.0	17.9
	Xwin-Math-7B	18.8	26.3	26.4	28.1	31.3	1.2	1.7
	Qwen2.5-math-Instruct-7b	34.4	**50.3**	34.4	**38.3**	**74.2**	**30.2**	**33.2**
Model > 7B	Abel-13B	31.2	48.0	37.2	22.4	24.4	0.8	0.8
	Xwin-Math-13B	29.8	45.9	33.1	30.2	31.3	1.2	1.7
	InternLM2-Math-Plus-20B	38.3	52.9	37.8	18.4	28.8	8.4	9.5
	QwQ-32B	**62.7**	**74.6**	**42.6**	**39.9**	**70.0**	**38.6**	**43.8**
Commercial models								
	OpenAI o1	68.9	78.6	**58.7**	60.1	55.8	39.8	40.9
	GPT-4o	38.1	48.3	37.7	59.0	44.7	19.7	21.3
	Deepseek-r1	**76.4**	**84.5**	55.6	**80.7**	86.8	54.2	65.3
	Claude3.7-sonnet	45.8	60.7	56.8	64.8	78.0	45.0	52.5
	Gemini2.5-pro	48.6	60.4	56.7	77.0	**90.8**	**65.1**	**75.7**

Internal difficulty-level analysis We further examine model performance across different difficulty levels within each field. In Algebra, models often experience sharp drops in accuracy on high-difficulty problems, indicating a gap in advanced atomic capabilities. In contrast, performance degradation in Geometry and Analysis is less severe. This suggests the need for training paradigms that better balance basic and advanced skill acquisition to ensure robust generalization. An interesting anomaly arises in Topology, where models sometimes perform better on harder problems than on easier ones. We hypothesize that this is due to mismatches between the models' training distributions and our evaluation data: some high-difficulty problems may incidentally align with abstract patterns the models have implicitly learned. This counterintuitive result shows that what humans find hard may not be hard for LLMs, which encourages deeper exploration into the field atomic ability decomposition.

Further results in Appendix C.2 show that training data difficulty significantly affects model performance. Notably, excessively low-difficulty training data may degrade accuracy across difficulty levels. Thus, balancing training data difficulty is essential for fostering generalizable atomic capabilities.

5.2 Experimental analysis of mathematical logical reasoning atomic capabilities

We evaluate several state-of-the-art mathematical models across multiple dimensions of logical reasoning. Detailed results are shown in Table 2. From these results, there are some insights that:

Models recognize definitions but struggle with deeper conceptual understanding All models perform better at recognizing definitions than completing missing properties, indicating a surface-level grasp of mathematical rigor. This suggests that while models can identify known concepts, they often lack precise internal representations. Larger commercial models significantly outperform smaller open-source ones; for example, `deepseek-r1` scores 84.5 in definition recognition and 76.4 in property completion. This disparity reflects the importance of pretraining, where larger models benefit from superior long-range memory and MoE (mixture-of-experts) mechanisms that mitigate knowledge forgetting. Current math models emphasize problem-solving over conceptual understanding, contributing to the gap between basic concept recognition and advanced reasoning or proof tasks. Section 5.4 provides further case studies.

Structured reasoning with formal language remains a challenge for smaller models Since large-scale commercial models show better performance with the best accuracy of 58.7, the models with smaller parameters struggle with both understanding the questions with formal math language and applying formal mathematical language to reason, despite performing reasonably well in natural-language-based reasoning. This suggests that intensive "problem-drilling" may promote pattern

Table 3: Performance comparison after stimulating various ***field*** atomic capabilities. We color the positive↑ / negative↓ influence as green / red.

Field	Algebra		Analysis		Geometry		Topology	
Difficulty Level	Low	High	Low	High	Low	High	Low	High
Qwen-base	80.5	65.2	67.7	66.5	52.1	53.4	52.1	53.4
Qwen-train-Algebra	80.2 (↓0.3)	69.7 (↑4.5)	75.8 (↑8.1)	71.5 (↑5.0)	65.7 (↑13.6)	57.5 (↑4.1)	56.0 (↑3.9)	62.3 (↑8.9)
Qwen-train-Analysis	81.3 (↑0.8)	66.4 (↑1.2)	71.8 (↑4.1)	64.9 (↓1.6)	58.2 (↑6.1)	55.8 (↑2.4)	46.1 (↓6.0)	54.1 (↑0.7)
Qwen-train-Geometry	79.6 (↓0.9)	68.1 (↑2.9)	69.8 (↑2.1)	59.5 (↓7.0)	57.3 (↑5.2)	56.0 (↑2.6)	52.8 (↑0.7)	60.9 (↑7.5)
Qwen-train-Topology	79.7 (↓0.8)	64.4 (↓0.8)	71.5 (↑3.8)	59.2 (↓7.3)	61.7 (↑9.6)	55.6 (↑2.2)	53.7 (↑1.6)	59.1 (↑5.7)

Table 4: Performance comparison after stimulating various ***field*** atomic capabilities that are trained on InternLM2-math-plus-7B.

Field	Algebra		Analysis		Geometry		Topology	
Difficulty Level	Low	High	Low	High	Low	High	Low	High
InternLM2-math-plus-7B	49.2	35.9	33.0	31.4	41.9	41.5	27.2	37.0
InternLM2-train-Algebra	48.1 (↓1.1)	37.8 (↑1.9)	35.8 (↑2.8)	33.6 (↑2.2)	43.1 (↑1.2)	42.1 (↑0.6)	27.8 (↑0.6)	37.9 (↑0.9)
InternLM2-train-Geometry	48.6 (↓0.6)	36.3 (↑0.4)	32.2 (↓0.8)	30.5 (↓0.8)	44.3 (↑2.4)	44.0 (↑2.5)	28.7 (↑1.5)	38.8 (↑1.8)

memorization over structured formal reasoning. Some models attempt to bridge this gap by generating code-like representations to aid multi-step deduction. Notably, reasoning-oriented models such as `o1` and `DeepSeek-R1`, equipped with stronger long-range inference and self-reflection, achieve outstanding results in this category.

Limited counterexample abilities reveal the limits of problem-solving training For the ability to judge the truth value of mathematical statements, open-source models average around 30 F-1 points, and even advanced models like `QWQ-32B` reach only 39.9, while `Deepseek-R1`, optimized for mathematical reasoning, scores 80.7. In generating counterexamples, the `Qwen` series performs particularly well, sometimes surpassing commercial models like `o1`. Conversely, models like `MetaMath` succeed in only 26.5% of such tasks. Although the best model Gemini2.5-pro demonstrates superior performance across various metrics when it constructs counter-examples, the performance of the other models on example consistency is still relatively low, with almost none exceeding 50%. This reflects the limitations of training paradigms overly focused on direct problem-solving, which hinders higher-level abstraction and conceptual reasoning.

5.3 Experimental analysis of interactions between atomic abilities

5.3.1 Influence between atomic abilities across different fields

We investigate the influence **between fields** among field atomic abilities, as summarized in Table 3. In particular, atomic abilities in *algebra* consistently exhibit **positive effects** across all other fields, with particularly strong gains observed in *Analysis* and *Geometry*. In some cases, Algebra training even yields greater improvements than in-field training. For instance, when evaluating *Geometry* atomic abilities, in-field training improved performance by 6.1 and 2.4 points at Levels 1 and 2, respectively, whereas activating Algebraic abilities led to larger gains of 13.6 and 4.2 points. This suggests a **complementary relationship** among atomic abilities, likely because algebraic problem-solving emphasizes fundamental reasoning skills that underpin more abstract fields, and our case study in Section 5.4 demonstrates that. These findings highlight the potential of leveraging **diverse, field-specific atomic abilities** to enhance target capabilities more effectively. However, we also observe **negative transfer** effects. Strengthening atomic abilities in *Topology* led to performance declines in *Algebra* and *Analysis*, with the largest drops reaching 0.8 and 7.3 points on Level 2. This may be due to substantial **data distribution divergence**. We have provided a visualization result in Appendix E. To further validate our conclusions, we have now included InternLM2-math-plus-7B as an additional baseline for training and evaluation, aiming to further strengthen our claims. The results, shown in Table 4, are consistent with our observations on Qwen2.5-math-instruct and provide a strong complement to our previous findings. These findings underscore the need to consider that,

Table 5: Performance comparison after stimulating various ***logic*** atom capabilities.

Atom Capability	Concept		Formal language	Counter example
	Attribute	Definition	Acc.	F-1
Qwen-base	34.4	50.3	34.4	30.2
Qwen-train-Concept	34.8 (↑0.4)	53.7 (↑3.4)	53.5 (↑19.1)	40.1 (↑9.9)
Qwen-train-Backward	30.3 (↓4.1)	46.3 (↓4.0)	50.3 (↑15.9)	41.1 (↑10.9)
Qwen-train-Forward	25.1 (↓9.3)	44.4 (↓5.9)	53.7 (↑19.3)	40.2 (↑10.0)

Table 6: Performance comparison after stimulating conceptual understanding of atomic capabilities.

Field	Algebra		Analysis		Geometry		Topology	
Difficulty Level	Low	High	Low	High	Low	High	Low	High
Qwen-base	80.5	65.2	67.7	66.5	52.1	53.4	52.1	53.4
Qwen-train-Concept	81.3 (↑0.8)	66.1 (↑0.9)	69.7 (↑2.0)	72.3 (↑5.8)	62.3 (↑10.2)	59.8 (↑6.4)	55.6 (↑3.5)	57.2 (↑3.8)

when trying to stimulate one, interactions between fields can significantly impact performance in unintended ways.

5.3.2 Interactions among logical reasoning atomic abilities

We also explore the interaction among **logical reasoning atomic abilities** as shown in Table 5. Our findings indicate that *conceptual understanding* plays a fundamental role in supporting two higher-level reasoning abilities. Training solely on fill-in-the-blank tasks that activate conceptual comprehension atomic ability results in substantial improvements of 19.1 and 9.9 in forward and backward reasoning. This aligns with our earlier analysis: current mathematical models, especially those with **limited parameter capacity**, exhibit significant deficiencies in conceptual understanding, which constrain the model's capacity to develop more advanced reasoning skills. Furthermore, forward reasoning and counterexample construction appear to **mutually reinforce** each other, indicating the potential for bidirectional enhancement between forward and backward reasoning atomic abilities. However, enhancing high-level atomic abilities alone can lead to a decline in conceptual understanding, revealing that **overemphasis on solving complex problems** may lead to **forgetting of basic mathematical concepts**.

5.3.3 Influence of conceptual understanding on field-specific atomic abilities

We also study the impact of logical atomic abilities on different fields. Particularly, we investigated how **conceptual understanding** supports field atomic abilities. The results in Table 6 show **significant performance improvements**, particularly in fields that rely on abstract reasoning, such as *Analysis and Geometry*. It is worth noting that activating conceptual understanding ability produces better outcomes than direct training on field-specific tasks. These results further confirm the **crucial role of conceptual understanding** as a foundational atomic ability, which highlights the need for future research to focus more on fostering **deep mathematical reasoning** and **concept-based learning**, rather than relying on **question repetition or difficulty escalation**.

5.4 Case study

To better understand the deficiencies of the model in some atomic capabilities and the impact of stimulating one atomic capability on another atomic capability, we conduct a case study. Specifically, we compare model predictions before and after stimulating algebra ability on a geometry task. As shown in Figure 3, before training, the model produces an incorrect result due to a flawed geometric assumption. After training, the model correctly applies knowledge of central angles and isosceles triangles, demonstrating the correlation from algebraic reasoning to geometric analysis.

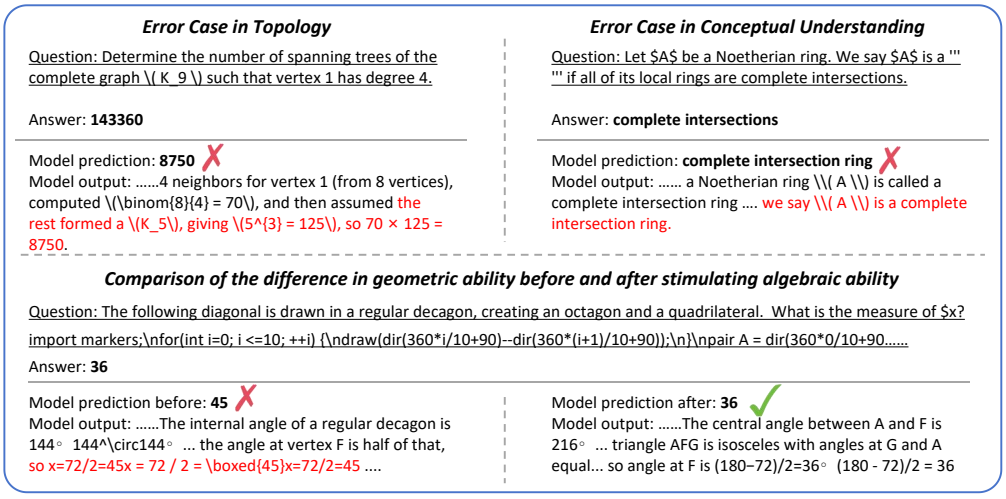

Figure 3: Case study about the error cases in Topology and Conceptual Atomic ability and the comparison of stimulating one atomic capability on another atomic capability.

6 Conclusion, limitation, and future directions

In our work, we focus on decoupling the mathematical reasoning ability of LLM into atomic abilities and exploring the interactions between atomic abilities. We design field atomic capabilities and logical reasoning atomic capabilities with data. The evaluation results and analysis of decoupled atomic abilities on advanced models highlight the limited performance of some abilities. We further explore the interactions, for which we observe whether other atomic capabilities are affected when one atomic capability is stimulated. This process provides very interesting inspirations, including the facilitating role of algebraic abilities and the supporting of conceptual comprehension ability, both on logical and field atomic abilities. However, we have not explored more advanced strategies to stimulate a specific atomic capability, such as curriculum learning or reinforcement learning. Our findings will encourage exploring how to better stimulate the required atomic abilities and utilize multiple atomic abilities to solve complex mathematical tasks.

Acknowledgement

This research is supported by Key-Area Research and Development Program of Guangdong Province, Granted No. 2024B1111060004. This research is also supported by Basic Research Fund of Shenzhen City (JCYJ20240813112009013).

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. Xing, *et al.*, "Lmsys-chat-1m: A large-scale real-world llm conversation dataset," in *The Twelfth International Conference on Learning Representations*.

[2] Y. Li, Q. Zhou, Y. Li, Z. Li, R. Liu, R. Sun, Z. Wang, C. Li, Y. Cao, and H. Zheng, "The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), pp. 3202–3213, Association for Computational Linguistics, 2022.

[3] X. Wu, Y.-L. Li, J. Sun, and C. Lu, "Symbol-llm: leverage language models for symbolic system in visual human activity reasoning," *Advances in neural information processing systems*, vol. 36, pp. 29680–29691, 2023.

[4] Y. Li, Q. Zhou, Y. Luo, S. Ma, Y. Li, H. Zheng, X. Hu, and P. S. Yu, "When llms meet cunning texts: A fallacy understanding benchmark for large language models," in *Advances in Neural*

Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[5] J. Du, Y. Wang, W. Zhao, Z. Deng, S. Liu, R. Lou, H. P. Zou, P. N. Venkit, N. Zhang, M. Srinath, H. Zhang, V. Gupta, Y. Li, T. Li, F. Wang, Q. Liu, T. Liu, P. Gao, C. Xia, C. Xing, C. Jiayang, Z. Wang, Y. Su, R. S. Shah, R. Guo, J. Gu, H. Li, K. Wei, Z. Wang, L. Cheng, S. Ranathunga, M. Fang, J. Fu, F. Liu, R. Huang, E. Blanco, Y. Cao, R. Zhang, P. S. Yu, and W. Yin, "Llms assist NLP researchers: Critique paper (meta-)reviewing," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024* (Y. Al-Onaizan, M. Bansal, and Y. Chen, eds.), pp. 5081–5099, Association for Computational Linguistics, 2024.

[6] Y. Li, S. Huang, X. Zhang, Q. Zhou, Y. Li, R. Liu, Y. Cao, H. Zheng, and Y. Shen, "Automatic context pattern generation for entity set expansion," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12458–12469, 2023.

[7] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 173:1–173:38, 2023.

[8] H. Wang, Y. Li, Y. Li, H. Zheng, W. Jiang, and H. Kim, "Exploring the implicit semantic ability of multimodal large language models: A pilot study on entity set expansion," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5, IEEE, 2025.

[9] J. Kuang, Y. Shen, J. Xie, H. Luo, Z. Xu, R. Li, Y. Li, X. Cheng, X. Lin, and Y. Han, "Natural language understanding and inference with MLLM in visual question answering: A survey," *ACM Comput. Surv.*, vol. 57, no. 8, pp. 190:1–190:36, 2025.

[10] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 2256–2264, 2024.

[11] J. Kuang, Y. Li, C. Wang, H. Luo, Y. Shen, and W. Jiang, "Express what you see: Can multimodal llms decode visual ciphers with intuitive semiosis comprehension?," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025* (W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds.), pp. 12743–12774, Association for Computational Linguistics, 2025.

[12] T. Zhou, D. Chen, Q. Jiao, B. Ding, Y. Li, and Y. Shen, "Humanvbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data," 2025.

[13] Z. Yue, L. Zhang, and Q. Jin, "Less is more: Mitigating multimodal hallucination from an eos decision perspective," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11766–11781, 2024.

[14] Q. Jiao, D. Chen, Y. Huang, B. Ding, Y. Li, and Y. Shen, "Img-diff: Contrastive data synthesis for multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9296–9307, 2025.

[15] Y. Li, Z. Xu, S. Chen, H. Huang, Y. Li, S. Ma, Y. Jiang, Z. Li, Q. Zhou, H. Zheng, and Y. Shen, "Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024* (L. Ku, A. Martins, and V. Srikumar, eds.), pp. 8656–8668, Association for Computational Linguistics, 2024.

[16] Y. Li, W. Zhang, Y. Yang, W. Huang, Y. Wu, J. Luo, Y. Bei, H. P. Zou, X. Luo, Y. Zhao, C. Chan, Y. Chen, Z. Deng, Y. Li, H. Zheng, D. Li, R. Jiang, M. Zhang, Y. Song, and P. S. Yu, "Towards agentic RAG with deep reasoning: A survey of rag-reasoning systems in llms," *CoRR*, vol. abs/2507.09477, 2025.

[17] A. Havrilla, S. Raparthy, C. Nalmpantis, J. Dwivedi-Yu, M. Zhuravynski, E. Hambro, and R. Raileanu, "Glore: when, where, and how to improve llm reasoning via global and local refinements," in *Proceedings of the 41st International Conference on Machine Learning*, pp. 17719–17733, 2024.

[18] X. Tan, H. Wang, X. Qiu, L. Cheng, Y. Cheng, W. Chu, Y. Xu, and Y. Qi, "Struct-x: Enhancing the reasoning capabilities of large language models in structured data scenarios," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2584–2595, 2025.

[19] L. Yuan, G. Cui, H. Wang, N. Ding, X. Wang, B. Shan, Z. Liu, J. Deng, H. Chen, R. Xie, *et al.*, "Advancing llm reasoning generalists with preference trees," in *The Thirteenth International Conference on Learning Representations*.

[20] X. Hu, X. Li, J. Chen, Y. Li, Y. Li, X. Li, Y. Wang, Q. Liu, L. Wen, P. S. Yu, and Z. Guo, "Evaluating robustness of generative search engine on adversarial factual questions," *CoRR*, vol. abs/2403.12077, 2024.

[21] Z. Yi, D. Zeng, Z. Ling, H. Luo, Z. Xu, W. Liu, J. Luan, W. Cao, and Y. Shen, "Attention Basin: Why Contextual Position Matters in Large Language Models," *arXiv e-prints*, p. arXiv:2508.05128, Aug. 2025.

[22] H. Luo, J. Kuang, W. Liu, Y. Shen, J. Luan, and Y. Deng, "Browsing like human: A multimodal web agent with experiential fast-and-slow thinking," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025* (W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds.), pp. 14232–14251, Association for Computational Linguistics, 2025.

[23] M. A. Islam, M. E. Ali, and M. R. Parvez, "Mapcoder: Multi-agent code generation for competitive problem solving," in *Annual Meeting of the Association of Computational Linguistics 2024*, pp. 4912–4944, Association for Computational Linguistics (ACL), 2024.

[24] J. Wang, M. Zerun, Y. Li, S. Zhang, C. Chen, K. Chen, and X. Le, "Gta: a benchmark for general tool agents," *Advances in Neural Information Processing Systems*, vol. 37, pp. 75749–75790, 2024.

[25] W. Zhang, Y. Bei, L. Yang, H. P. Zou, P. Zhou, A. Liu, Y. Li, H. Chen, J. Wang, Y. Wang, F. Huang, S. Zhou, J. Bu, A. Lin, J. Caverlee, F. Karray, I. King, and P. S. Yu, "Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap," *CoRR*, vol. abs/2501.01945, 2025.

[26] X. Tan, S. Shi, X. Qiu, C. Qu, Z. Qi, Y. Xu, and Y. Qi, "Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness," in *Proceedings of the 2023 conference on empirical methods in natural language processing: industry track*, pp. 650–662, 2023.

[27] G. Chujie, S. Wu, Y. Huang, D. Chen, Q. Zhang, Z. Fu, Y. Wan, L. Sun, and X. Zhang, "Honestllm: Toward an honest and helpful large language model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 7213–7255, 2024.

[28] W. Hua, X. Yang, M. Jin, Z. Li, W. Cheng, R. Tang, and Y. Zhang, "Trustagent: Towards safe and trustworthy llm-based agents through agent constitution," in *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.

[29] Y. Li, H. Huang, J. Kuang, Y. Li, S. Guo, C. Qu, X. Tan, H. Zheng, Y. Shen, and P. S. Yu, "Refine knowledge of large language models via adaptive contrastive learning," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, 2025.

[30] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[31] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[32] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang, "Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct," *arXiv preprint arXiv:2308.09583*, 2023.

[33] L. Yu, W. Jiang, H. Shi, Y. Jincheng, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu, "Metamath: Bootstrap your own mathematical questions for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[34] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[35] H. Ying, S. Zhang, L. Li, Z. Zhou, Y. Shao, Z. Fei, Y. Ma, J. Hong, K. Liu, Z. Wang, *et al.*, "Internlm-math: Open math large language models toward verifiable reasoning," *arXiv preprint arXiv:2402.06332*, 2024.

[36] E. Chern, H. Zou, X. Li, J. Hu, K. Feng, J. Li, and P. Liu, "Generative ai for math: Abel." `https://github.com/GAIR-NLP/abel`, 2023.

[37] OpenAI, "Gpt-4 technical report," 2023.

[38] Y. Li, Y. Li, X. Wang, Y. Jiang, Z. Zhang, X. Zheng, H. Wang, H. Zheng, F. Huang, J. Zhou, and P. S. Yu, "Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, 2025.

[39] Z. Xu, Y. Li, R. Ding, X. Wang, B. Chen, Y. Jiang, H. Zheng, W. Lu, P. Xie, and F. Huang, "Let llms take on the latest challenges! A chinese dynamic question answering benchmark," in *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025* (O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, eds.), pp. 10435–10448, Association for Computational Linguistics, 2025.

[40] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu, "Metamath: Bootstrap your own mathematical questions for large language models," *arXiv preprint arXiv:2309.12284*, 2023.

[41] C. Li, W. Wang, J. Hu, Y. Wei, N. Zheng, H. Hu, Z. Zhang, and H. Peng, "Common 7b language models already possess strong math capabilities," *arXiv preprint arXiv:2403.04706*, 2024.

[42] L. d. Moura and S. Ullrich, "The lean 4 theorem prover and programming language," in *Automated Deduction – CADE 28* (A. Platzer and G. Sutcliffe, eds.), (Cham), pp. 625–635, Springer International Publishing, 2021.

[43] H. Lin, Z. Sun, Y. Yang, and S. Welleck, "Lean-star: Learning to interleave thinking and proving," *arXiv preprint arXiv:2407.10040*, 2024.

[44] W. Sun, Q. Du, F. Cui, and J. Zhang, "An efficient and precise training data construction framework for process-supervised reward model in mathematical reasoning," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025* (W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds.), pp. 4292–4305, Association for Computational Linguistics, 2025.

[45] X. Tan, T. Yao, C. Qu, B. Li, M. Yang, D. Lu, H. Wang, X. Qiu, W. Chu, Y. Xu, *et al.*, "Aurora: Automated training framework of universal process reward models via ensemble prompting and reverse verification," *arXiv preprint arXiv:2502.11520*, 2025.

[46] A. Didolkar, A. Goyal, N. R. Ke, S. Guo, M. Valko, T. P. Lillicrap, D. J. Rezende, Y. Bengio, M. C. Mozer, and S. Arora, "Metacognitive capabilities of llms: An exploration in mathematical problem solving," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada,*

December 10 - 15, 2024 (A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, eds.), 2024.

[47] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024.

[48] X. Yu, B. Zhou, H. Cheng, and D. Roth, "Reasonagain: Using extractable symbolic programs to evaluate mathematical reasoning," *arXiv preprint arXiv:2410.19056*, 2024.

[49] H. Luo, Y. Deng, Y. Shen, S. K. Ng, and T.-S. Chua, "Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7978–7993, 2024.

[50] F. Teng, Z. Yu, Q. Shi, J. Zhang, C. Wu, and Y. Luo, "Atom of thoughts for markov llm test-time scaling," *arXiv preprint arXiv:2502.12018*, 2025.

[51] Z. Ling, D. Chen, L. Yao, Q. Shen, Y. Li, and Y. Shen, "Diversity as a reward: Fine-tuning llms on a mixture of domain-undetermined data," *arXiv preprint arXiv:2502.04380*, 2025.

[52] S. Huang, S. Ma, Y. Li, M. Huang, W. Zou, W. Zhang, and H. Zheng, "Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy* (N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds.), pp. 10186–10197, ELRA and ICCL, 2024.

[53] T. Yu, C. Jiang, C. Lou, S. Huang, X. Wang, W. Liu, J. Cai, Y. Li, Y. Li, K. Tu, H. Zheng, N. Zhang, P. Xie, F. Huang, and Y. Jiang, "Seqgpt: An out-of-the-box large language model for open domain sequence understanding," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada* (M. J. Wooldridge, J. G. Dy, and S. Natarajan, eds.), pp. 19458–19467, AAAI Press, 2024.

[54] A. Amini, S. Gabriel, P. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, "Mathqa: Towards interpretable math word problem solving with operation-based formalisms," *arXiv preprint arXiv:1905.13319*, 2019.

[55] H. Liu, Z. Zheng, Y. Qiao, H. Duan, Z. Fei, F. Zhou, W. Zhang, S. Zhang, D. Lin, and K. Chen, "Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark," 2024.

[56] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," in *International Conference on Learning Representations (ICLR)*, 2024.

[57] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, and P. S. Yu, "Large language models meet NLP: A survey," *CoRR*, vol. abs/2405.12819, 2024.

[58] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, and P. S. Yu, "Multilingual large language model: A survey of resources, taxonomy and frontiers," *CoRR*, vol. abs/2404.04925, 2024.

[59] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[60] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen, *et al.*, "Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions," *Hugging Face repository*, vol. 13, p. 9, 2024.

[61] Y. Li, H. Huang, S. Ma, Y. Jiang, Y. Li, F. Zhou, H. Zheng, and Q. Zhou, "On the (in)effectiveness of large language models for chinese text correction," *CoRR*, vol. abs/2307.09007, 2023.

[62] Y. Li, S. Qin, J. Ye, S. Ma, Y. Li, L. Qin, X. Hu, W. Jiang, H. Zheng, and P. S. Yu, "Rethinking the roles of large language models in chinese grammatical error correction," *CoRR*, vol. abs/2402.11420, 2024.

[63] K. Yang and J. Deng, "Learning to prove theorems via interacting with proof assistants," in *International Conference on Machine Learning*, pp. 6984–6994, PMLR, 2019.

[64] K. Zheng, J. M. Han, and S. Polu, "minif2f: a cross-system benchmark for formal olympiad-level mathematics," in *International Conference on Learning Representations*, 2022.

[65] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *NeurIPS*, 2021.

[66] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[67] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun, "Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems," 2024.

[68] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, and S. Chaudhuri, "Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[69] X. Zhang, C. Li, Y. Zong, Z. Ying, L. He, and X. Qiu, "Evaluating the performance of large language models on gaokao benchmark," *arXiv preprint arXiv:2305.12474*, 2023.

[70] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[71] Z. He, T. Liang, J. Xu, Q. Liu, X. Chen, Y. Wang, L. Song, D. Yu, Z. Liang, W. Wang, *et al.*, "Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning," *arXiv preprint arXiv:2504.11456*, 2025.

[72] S. Welleck, J. Liu, R. L. Bras, H. Hajishirzi, Y. Choi, and K. Cho, "Naturalproofs: Mathematical theorem proving in natural language," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual* (J. Vanschoren and S. Yeung, eds.), 2021.

[73] Y. Li, J. Kuang, H. Huang, Z. Xu, X. Liang, Y. Yu, W. Lu, Y. Li, X. Tan, C. Qu, *et al.*, "One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical llms," *arXiv preprint arXiv:2502.10454*, 2025.

[74] G. Cui, L. Yuan, Z. Wang, H. Wang, W. Li, B. He, Y. Fan, T. Yu, Q. Xu, W. Chen, J. Yuan, H. Chen, K. Zhang, X. Lv, S. Wang, Y. Yao, H. Peng, Y. Cheng, Z. Liu, M. Sun, B. Zhou, and N. Ding, "Process reinforcement through implicit rewards," 2025.

[75] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims well reflect the motivation and contribution of this paper and provides inspiration for subsequent research.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and future directions in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work mainly explores the atomic capabilities of LLM mathematics reasoning from an experimental perspective as an empirical study, and does not conduct theoretical research.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the models, hardware information, prompt design, etc. used for training and evaluation in Section 4 and the Appendix B, and provide our data and code in the supplementary materials for review. We will further disclose the relevant data and code to facilitate subsequent researchers to reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our data and code in the supplementary materials for review. We will further disclose the relevant data and code in GitHub and Huggingface to facilitate subsequent researchers to reproduce our results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the models, hardware information, prompt design, etc. used for training and evaluation in Section 4 and the Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the limitation of experimental resources (especially for expensive commercial models) and considering that these studies cannot support the main claims, we did not study the error bars, confidence intervals and other statistical data in our experiments. In addition, we have provided sufficient experimental details, data, and codes to ensure the reproducibility and credibility of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details of the compute resources used to conduct the experiments are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We promise that we strictly follow the code of ethics of NeurIPS 2025.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We describe our impact in the Section 1, 4, and 6, and declare that our research has no conflicts of interest and will not cause negative impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We have not released models and data with high risk of abuse. All the models and data used in our research have been strictly risk-controlled and open-source.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The models and data we use are all official open source, or call the official API. All use complies with their license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All the data and codes we processed are organized and documented, and submitted together with the supplementary materials for review. We will further open source these data and codes to allow more researchers to participate in the exploration.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in our research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional details about data

After our data construction process, we collected the data corresponding to each decoupled atomic capability and divided it into training set and test set. The specific statistical results are as following Table 7 and Table 8:

Table 7: Data statistics of **FIELD** atomic capabilities.

Field Cap.	Algebra		Geometry		Analysis		Topology	
	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2
Train/Test	3813/1277	4517/1505	3351/1117	3391/1331	3276/1092	4077/1358	3336/1112	3176/1058

Table 8: Data statistics of **LOGICAL** atomic capabilities.

Logic Cap.	Conceptual Understanding		Forward Reasoning	Backward Reasoning
	Attribute Description	Definition	Formal Language	Counter-example
Train/Test	1225/1217	1683/1661	1061/1032	1225/1217

In addition, we provide data examples of our logical atomic capabilities as shown in Figure 4, to help readers better understand the different logical reasoning atomic tasks.

Attribute Description Task in Conceptual Understanding Atmoic Abilities

"**title**": "fields-definition-linearly-disjoint"

"**question**": "Consider a diagram of fields as in equation-inside-omega. We say that K and L are ''' ''' if the map $$ K \\otimes_k L \\longrightarrow KL,\\quad \\sum x_i \\otimes y_i \\longmapsto \\sum x_i y_i $$ is injective.",

"**final_answer**": ["**linearly disjoint over k in $\\Omega$**"]}

Definition Task in Conceptual Understanding Atmoic Abilities

"**title**": "Definition:Reciprocal",

"**question**": "Let $x \\in \\R$ be a real number such that $x \\ne 0$. Then $\\dfrac 1 x$ is called the ''' '''."

"**final_answer**": ["**reciprocal of x**"]

Formal Language Multi-step Reasoning in Forward Reasoning Atmoic Abilities

"**question**": " (a : ℕ → ℝ) (a0 : a 0 = 1) (a1 : a 1 = 2) (a_rec : ∀ n, a (n + 2) = 4 * a (n + 1) + a n) : ∃ f : ℕ → ℝ , ∀ n, a n = f n := by sorry",

"**final_answer**": "The closed form is $a_n=\\frac{1}{2}((2-\\sqrt{5})^n+(2+\\sqrt{5})^n)$"

"**solution**": [.....use a\nintros n\nsimp [a_rec, a0, a1]", "use a\nsimp [a0, a1, a_rec]", "use fun n ↦ a n\nintro n\ninduction' n with n IH\nsimp [a0]\nsimp [a_rec, IH]", "refine' ⟨ a, fun n => _⟩ \ninduction n <;> simp [a_rec, a0, a1, *]", "refine ⟨ a,?_⟩ \nintro n\ninduction n <;> simp [a_rec, *]", "use a\nintro n\ninduction n <;> simp [a_rec, *]", "use fun n => a n\nintro n\ninduction n <;> simp [a0, a1, a_rec, *]", "refine' ⟨ a, _⟩ \nintro n\ninduction' n with n IH\nsimp [a0]\nsimp [a1, a_rec, IH]", "use fun n ↦ a n\nintro n\nrfl", "let f : ℕ → ℝ := fun n ↦ a n\nrefine' ⟨ f, fun n ↦ rfl⟩ .....

Counter-example-driven Reverse Reasoning Atmoic Abilities

"**question**": "a^{m} is not necessarily the smallest positive power of square of $\\left(a^{m}\\right)$ in terms of a.",

"**final_answer**": "True",

"**solution**": "For example, in the 6-element cyclic group $G=(a)$, we have $(a)=\\left(a^{5}\\right)$."

Figure 4: The figure presents data examples in logical atomic capability, including property descriptions, definition recognition, formal mathematical language-driven forward reasoning, and counterexample-driven backward reasoning.

B Additional details about the experimental prompts and metrics

Prompt used for experiments The used prompts are summarized as follows. For field atomic capability, we use the same default chain of thought prompt:

Prompt for Field Atomic Capability

Please think step by step to solve the following question, and put your final answer within \\boxed{}.
{question}

For conceptual understanding atomic capability, wo prompt the LLMs to fill in the blank:

Prompt for conceptual understanding Capability

Please think step by step to fill in the blank in ”’ ”’ of following statement, and put your final answer within \\boxed{}.
{question}

For forward multi-step reasoning, we prompt the model to use formal math language to solve the question:

Prompt for forward reasoning capability

Please think step by step to solve the following question by formal math language, and put your final answer within \\boxed{}.
{question}

For counter-example-driven backward reasoning, we prompt the model to judge the statement true or false, which is the same setting of counter-math:

Counter-example backward reasoning prompt

{statement}
Please think step by step about whether the above statement is True or False, and put your final answer within \\boxed{}.

Evaluation metrics of counter-example based backward reasoning In the evaluation of counterexample-driven backward reasoning, we go beyond computing the F1 score for judging the truthfulness of a given statement. Inspired by the CounterMATH framework, we also assess GPT-4o's capability to generate examples within its reasoning output. This is achieved through Example Extraction, which detects and retrieves instances where the model explicitly introduces or references counterexamples to support its claims. Alignment Assessment then determines whether each extracted example is consistent with a predefined Reference Example in terms of logical reasoning pattern, problem decomposition strategy, and goal relevance. Notably, since a proposition may have multiple valid counterexamples, exact replication of the reference is not mandatory for determining consistency. Instead, the reference serves as a guiding benchmark for GPT-4o, mitigating the risk of fully autonomous evaluations that may diverge from human-aligned reasoning standards.

Specifically, the Examples metric indicates the percentage of problems where the model incorporates examples in its solution process. Strict Align represents the fraction of model-generated examples that fully match the reference in reasoning alignment, while Loose Align captures the proportion of cases where at least one example aligns with the reference example.

C Additional experiment and analysis of impact of atomic abilities across difficulty levels within the same field

C.1 Additional Results in Field Interaction

We run each experiment multiple times with different seeds and report means and standard deviations as shown in Table 9.

Table 9: Performance comparison across various mathematical fields.

Field	Algebra		Analysis		Geometry		Topology	
Level	Low	High	Low	High	Low	High	Low	High
Qwen-base	80.5	65.2	67.7	66.5	52.1	53.4	52.1	53.4
Qwen-train-Algebra	80.2 (±0.4)	69.7 (±0.5)	75.8 (±0.6)	71.5 (±0.7)	65.7 (±0.3)	57.5 (±0.6)	56.0 (±0.5)	62.3 (±0.9)
Qwen-train-Geometry	79.6 (±0.5)	68.1 (±0.4)	69.8 (±0.6)	59.5 (±0.8)	57.3 (±0.6)	56.0 (±0.5)	52.8 (±0.4)	60.9 (±0.3)

C.2 Difficulty Interaction in Field Atomic Ability

We conduct comparative experiments within each mathematical field to investigate how training on datasets of varying difficulty levels affects the development of atomic abilities. The results in Table 11 demonstrate that models trained on **high-difficulty data** exhibit performance improvements on both easy and hard test, with particularly notable gains on the easier tasks. This indicates that mastering complex knowledge not only activates high-level atomic abilities but also facilitates the co-activation of lower-level atomic abilities. However, training solely on low-difficulty data can negatively affect performance on harder tasks. This phenomenon is especially evident in fields such as *Algebra, Analysis, and Topology*, where models trained on easier data show a performance decline when evaluated on more difficult tasks. These findings suggest the need for careful **balancing of difficulty distribution** in training datasets to stimulate atomic abilities across the full spectrum and prevent **knowledge forgetting**.

Table 10: Performance comparison of different training levels across various mathematical fields. We color the positive↑ / negative↓ influence as green / red.

Field	Algebra		Analysis		Geometry		Topology	
Level	Low	High	Low	High	Low	High	Low	High
Qwen-base	80.5	65.2	67.7	66.5	52.1	53.4	52.1	53.4
Qwen-train-l2	83.2 (↑2.7)	68.3 (↑3.1)	76.8 (↑9.1)	66.4 (↓0.1)	59.6 (↑7.5)	57.7 (↑4.3)	53.1 (↑1.0)	53.8 (↑0.4)
Qwen-train-l1	79.5 (↓1.0)	64.9 (↓0.3)	70.5 (↑2.8)	62.2 (↓4.3)	58.9 (↑6.8)	57 (↑3.6)	52.2 (↑0.1)	52.1 (↓1.3)
Qwen-train-all	80 (↓0.5)	69.7 (↑4.5)	71.8 (↑4.1)	64.9 (↓1.6)	57.4 (↑5.3)	56.1 (↑2.7)	53.7 (↑1.6)	59.1 (↑5.7)

Table 11: Performance comparison of different training levels across various mathematical fields. We color the positive↑ / negative↓ influence as green / red.

Field	Algebra		Analysis		Geometry		Topology	
Level	Low	High	Low	High	Low	High	Low	High
Qwen-base	80.5	65.2	67.7	66.5	52.1	53.4	52.1	53.4
Qwen-train-l2	83.2 (↑2.7)	68.3 (↑3.1)	76.8 (↑9.1)	66.4 (↓0.1)	59.6 (↑7.5)	57.7 (↑4.3)	53.1 (↑1.0)	53.8 (↑0.4)
Qwen-train-l1	79.5 (↓1.0)	64.9 (↓0.3)	70.5 (↑2.8)	62.2 (↓4.3)	58.9 (↑6.8)	57 (↑3.6)	52.2 (↑0.1)	52.1 (↓1.3)
Qwen-train-all	80 (↓0.5)	69.7 (↑4.5)	71.8 (↑4.1)	64.9 (↓1.6)	57.4 (↑5.3)	56.1 (↑2.7)	53.7 (↑1.6)	59.1 (↑5.7)

D Human Evaluation

We conduct a human evaluation on a sample of the data during the initial phase of our experiments in order to verify the accuracy of our automatic evaluation metrics. To do this, we selected two models,

Gemini and Qwen, and sampled 20 problems from each dataset for manual review. Furthermore, since the majority of the problems are calculation-based, the final answers have a relatively fixed format. This, combined with our constraint that the answer must be in the
boxed format, greatly simplified the validation process and made it easier to check for accuracy. The table below shows the number of cases (out of 20) where the automatic evaluation aligned with the human evaluation for each model and dataset.

Table 12: The human evaluation category-wise scores.

Model	**Algebra**		**Geometry**		**Analysis**		**Topology**		**Attr.**	**Def.**	**Forward Rea.**
	l1	l2	l1	l2	l1	l2	l1	l2			
Qwen2.5-Math-Instruct-7B	20/20	20/20	20/20	20/20	20/20	20/20	19/20	19/20	18/20	20/20	20/20
Gemini2.5-pro	20/20	20/20	20/20	20/20	19/20	20/20	20/20	20/20	19/20	20/20	20/20

E Visualization results of field atomic ability interaction

We have provided a heat map in Figure 5 that represents the correlation between atomic capabilities in the field, providing a more intuitive result.

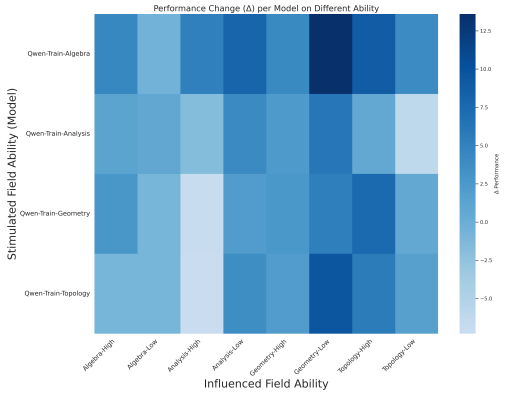

Figure 5: This figure shows the impact of stimulating one ability on the remaining abilities, where deeper colors indicate a greater positive facilitation effect, while lighter colors indicate a greater negative impact.