A Survey of Mathematical Reasoning in the Era of Multimodal Large Language Model: Benchmark, Method & Challenges

Anonymous ACL submission

Abstract

Mathematical reasoning, a core aspect of human cognition, is vital across many domains, from educational problem-solving to scientific advancements. As artificial general intelligence (AGI) progresses, integrating large language models (LLMs) with mathematical reasoning tasks is becoming increasingly significant. This survey provides the **first comprehensive analysis of mathematical reasoning in the era of multimodal large language models (MLLMs)**. We review over 200 studies published since 2021, and examine the state-of-the-art developments in Math-LLMs, with a focus on multimodal settings. We categorize the field into three dimensions: *benchmarks*, *methodologies*, and *challenges*. In particular, we explore multimodal mathematical reasoning pipeline, as well as the role of (M)LLMs and the associated methodologies. Finally, we identify seven major challenges hindering the realization of AGI in this domain, offering insights into the future direction for enhancing multimodal reasoning capabilities. This survey serves as a critical resource for the research community in advancing the capabilities of LLMs to tackle complex multimodal reasoning tasks.

1 Introduction

Mathematical reasoning is a critical aspect of human cognitive ability, involving the process of deriving conclusions from a set of premises through logical and systematic thinking (Jonsson et al., 2022; Yu et al., 2024b). It plays an essential role in a wide range of applications, from problem-solving in education to advanced scientific discoveries. As artificial general intelligence (AGI) continues to advance (Zhong et al., 2024), the integration of large language models (LLMs) with mathematical reasoning tasks becomes increasingly significant. These models, with their impressive capabilities in language understanding, have the potential to simulate complex reasoning processes that were once thought to be inherently human. In recent years, both academia and industry have placed increasing emphasis on this direction (Wang et al., 2024d; Xu et al., 2024a; Lu et al., 2022b; Yan et al., 2025).

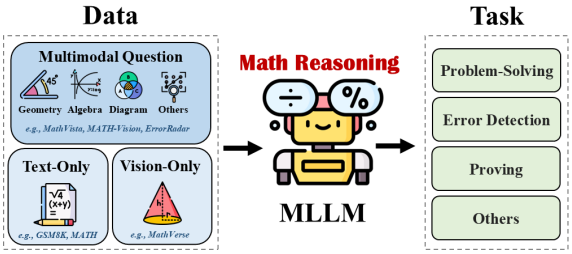

Figure 1: The illustration of our research scope (*i.e.,* investigating the MLLM's math reasoning capability).

Survey	Venue & Year	Scope	Multimodal	LLM
(O'Halloran, 2015)	JMB'15	MM4Math	✔	
(Hegedus and Tall, 2015)	IRME'15	MM4Math	✔	
(Lu et al., 2022b)	ACL'22	DL4Math		
(Li et al., 2023a)	arXiv'23	LLM4Edu		✔
(Liu et al., 2023b)	arXiv'23	LLM4Edu		✔
(Li et al., 2024g)	COLM'24	DL4TP		
(Ahn et al., 2024)	EACL'24	LLM4Math		✔
(Xu et al., 2024a)	IJMLC'24	LLM4Edu		✔
(Wang et al., 2024d)	arXiv'24	LLM4Edu		✔
Ours	-	MLLM4Math	✔	✔

Table 1: Comparisons between relevant surveys & ours.

The inputs for mathematical reasoning tasks are diverse, extending beyond traditional text-only to multimodal settings, as illustrated in Figure 1. Mathematical problems often involve not only textual information but also visual elements, such as diagrams, graphs, or equations, which provide essential context for solving the problem (Wang et al., 2024e; Yin et al., 2024). In the past year, multimodal mathematical reasoning has emerged as a key focus for multimodal large language models (MLLMs) (Zhang et al., 2024c; Bai et al., 2024; Wu et al., 2023a). This shift is driven by the recognition that reasoning tasks in fields like mathematics require models capable of integrating and processing multiple modalities simultaneously to achieve human-like performance. However, multimodal mathematical reasoning poses significant

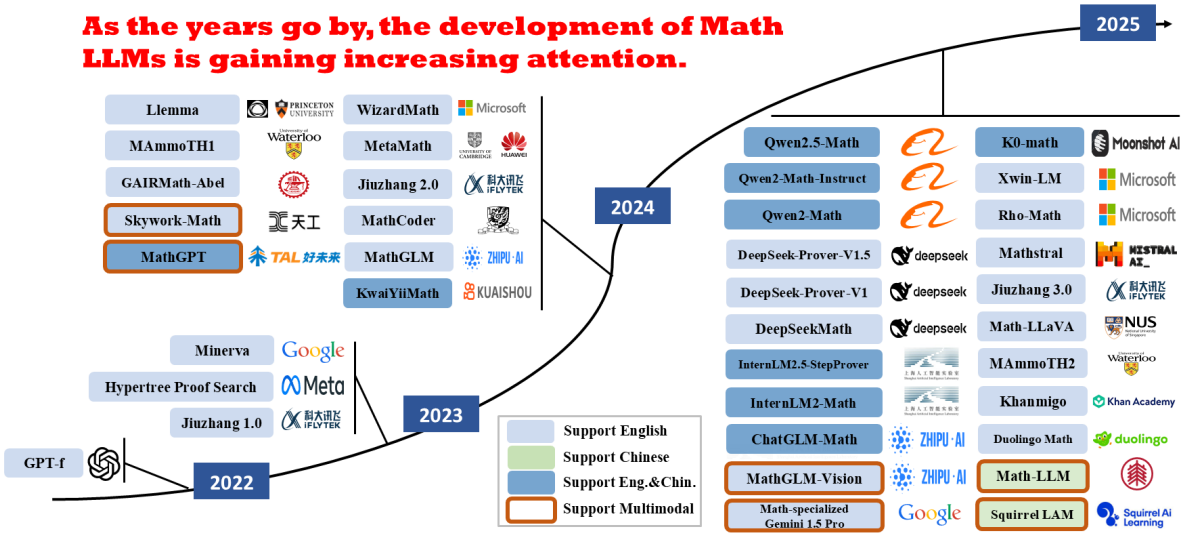

Figure 2: The release timeline of Math-LLMs in recent years.

challenges due to the complex interaction between different modalities, the need for deep semantic understanding, and the importance of context preservation across modalities (Liang et al., 2024a; Song et al., 2023; Fu et al., 2024b). These challenges are central to the realization of AGI, where models must integrate diverse forms of knowledge seamlessly to perform sophisticated reasoning tasks.

Math-LLM Progress. Figure 2 illustrates that, driven by the rapid development of LLMs since 2021, the number of math-specific LLMs (Math-LLMs) has grown steadily, alongside enhanced support for multilingual and multimodal capabilities (More details in Appendix A). The landscape was marked by the introduction of models like GPT-f (Polu and Sutskever, 2021) and Minerva (Lewkowycz et al., 2022), with Hypertree Proof Search (Lample et al., 2022) and Jiuzhang 1.0 (Zhao et al., 2022) highlighting advancements in theorem proving and mathematical question understanding capabilities, respectively. Year 2023 saw a surge in diversity and specialization, alongside multimodal support from models like Skywork-Math (Zeng et al., 2024). In year 2024, there was a clear focus on enhancing mathematical instruction (*e.g.*, Qwen2.5-Math (Yang et al., 2024a)) and proof (*e.g.*, DeepSeek-Proof (Xin et al., 2024a)) capabilities. The year also witnessed the emergence of Math-LLMs with a vision component, such as MathGLM-Vision (Yang et al., 2024b).

Scope. Previous surveys have not fully captured the progress and challenges of mathematical reasoning in the age of MLLMs. As indicated in Table 1, some works have concentrated on the application of deep learning techniques to mathematical reasoning (Lu et al., 2022b) or specific domains such as theorem proving (Li et al., 2024g), but they have overlooked the rapid advancements brought about by the rise of LLMs. Others have broadened the scope to include the role of LLMs in education (Wang et al., 2024d; Xu et al., 2024a; Li et al., 2023a) or mathematical fields (Ahn et al., 2024; Liu et al., 2023b), but have failed to explore the development and challenges of mathematical reasoning in multimodal settings in depth. Therefore, this survey aims to fill this gap by providing the **first-ever comprehensive analysis of the current state of mathematical reasoning in the era of MLLMs**, focusing on three key dimensions: *benchmark*, *methodology*, and *challenges*.

Structure. In this paper, we survey over 200 publications from the AI community since 2021 related to (M)LLM-based mathematical reasoning, and summarize the progress of Math-LLMs. We first approach the field from the benchmark perspective, analyzing the LLM-based mathematical reasoning task through four key aspects: basic focus, task, evaluation, and training data (Section 2). Subsequently, we explore the roles that (M)LLMs play in mathematical reasoning, categorizing them as reasoners, enhancers, and planners (Section 3). Finally, we identify seven core challenges that the mathematical reasoning faces in the era of MLLMs (Section 4). This survey aims to provide the community with comprehensive insights for advancing multimodal reasoning capabilities of LLMs.

2 Benchmark Perspective

2.1 Overview

Benchmarking for mathematical reasoning plays a crucial role in advancing LLM research, as it provides standardized, reproducible pipeline for assessing the performance on reasoning tasks. While previous benchmarks such as GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019) were instrumental in the pre-LLM era, our scope is centered on those relevant to (M)LLMs. In this section, we present a comprehensive analysis of recent benchmarks for mathematical reasoning in the context of (M)LLMs (Shown in Table 3 from Appendix B). The section is organized into four subsections: Basic Focus (Sec.2.2), Tasks (Sec.2.3), Evaluation (Sec.2.4), and Training Data (Sec.2.5).

2.2 Basic Focus

Basic Format. In a math reasoning task (taking prblem-solving as a basic setting), the goal is to solve a mathematical problem given a specific format of input and output. The input consists of a statement that describes the problem to be solved. As shown in Figure 3, this can be presented in either a textual format or a multimodal format (text accompanied by visual elements, such as figures or diagrams). The output is the predicted solution to the problem, represented as numerical or symbolic results. More cases can be seen in Appendix C.

Language & Size. The majority of benchmarks are available in English, with a few exceptions like Chinese (Li et al., 2024i) or Romanian (Cosma et al., 2024) datasets. This predominance of English datasets underscores the challenges of multilingual representation in the mathematical reasoning domain, suggesting an opportunity for future work to diversify datasets across languages, especially those in underrepresented regions. Moreover, the size of these datasets varies widely, from smaller sets (*e.g.*, QRData (Liu et al., 2024d) with 411 questions) to massive corpora (*e.g.*, OpenMathInstruct-1 (Toshniwal et al., 2024b) with 1.8 million problem-solution pairs). Larger datasets are more likely to support robust model training and evaluation, but their size can also present challenges in terms of computational requirements and quality control.

Source. The sources of datasets predominantly consist of public (*i.e.*, derived from public repositories or datasets) and private sources. The private datasets typically offer specialized problem types and tasks, and may present unique challenges, such as restricted access or ethical considerations. On the other hand, public datasets foster wider community collaboration, though they may suffer from limitations in diversity and task coverage. Some works have also leveraged LLMs to generate the datasets tailored to specific needs. For instance, GeomVerse constructs synthetic datasets to evaluate the multi-hop reasoning abilities required in geometric math problems (Kazemi et al., 2023).

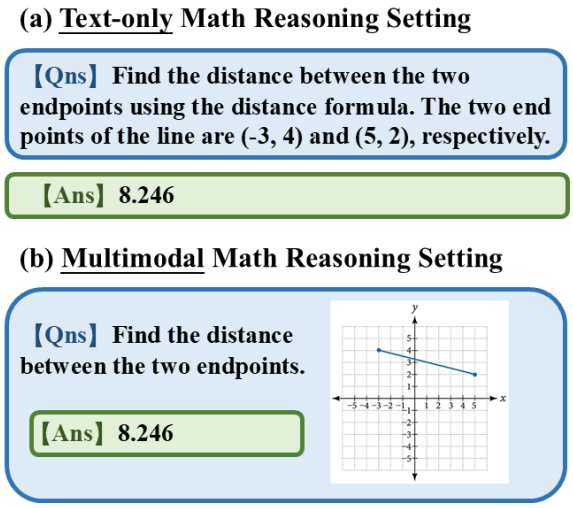

Figure 3: Typical data format of math reasoning task for text-only & multimodal settings. Examples are derived from MathVerse (Zhang et al., 2024f), which assess whether and how much MLLMs can truly understand the visual diagrams for mathematical reasoning.

Educational Level. The benchmarks span various educational levels, ranging from elementary school to university-level problems. Besides, there has also been a surge in datasets focused on competition-level problems (Tsoukalas et al.), offering insights into the current limitations of LLMs in comparison to the upper bound of human cognitive abilities. Future directions could involve more focused datasets targeting specific educational levels to enable models to specialize in handling particular age groups or skill sets.

2.3 Task

Model Choice. The choice of models in these benchmarks spans open-source and closed-source models, with a growing interest in Math-LLMs. This trend indicates an increasing recognition of the need for models tailored to mathematical reasoning, which often require specialized training and handling of structured knowledge. Additionally, with the recent release of GPT-4o (OpenAI, 2024)

and Gemini-Pro-1.5 (Reid et al., 2024), which have demonstrated significant advancements in multimodal reasoning capabilities, the latest benchmarks have begun to include them in the evaluations. For example, ErrorRadar, in its initial formulation of multimodal error detection setting, incorporates these state-of-the-art MLLMs to highlight the real-world performance gap between AI systems and human-level reasoning (Yan et al., 2024a).

Reasoning Task. Problem-solving tasks typically dominate, reflecting the emphasis on students' ability to apply knowledge and reasoning skills in real-world contexts. This also serves as the core objective of current Math-LLMs. In addition, a growing proportion of error detection tasks suggests an increasing focus on helping students recognize and correct mistakes (Li et al., 2024e; Yan et al., 2024a; Kurtic et al., 2024). Meanwhile, proving tasks, often associated with higher-order thinking, highlight a shift towards cultivating logical reasoning and systematic problem-solving abilities (Tsoukalas et al.). Moreover, a smaller portion of work has addressed tasks that align with real-world educational needs but lack systematic formulation. For instance, Li et al. (2024e) further introduces error correction (which goes beyond simple error detection); Didolkar et al. (2024) explores automated skill discovery for problem-solving; and MathChat (Liang et al., 2024c) focuses on reasoning in multi-turn settings (such as follow-up QA and problem generation). Given the higher demands on reasoning capabilities in multimodal settings, many studies have also evaluated the aforementioned reasoning tasks in image-text problem settings. These efforts aim to provide the LLM community with more diverse, real-world task scenarios, catering to the needs of multimodal learning environments.

2.4 Evaluation

Discriminative Evaluation is a common approach, focusing on the ability of M(LLM)s to correctly classify or choose the correct answer (Hendrycks et al., 2021; Mishra et al., 2022; Li et al., 2024c). Based on specific motivations, some works also build their metrics upon accuracy for further expansion. For example, GSM-PLUS, a new adversarial benchmark for evaluating the robustness of LLMs in mathematical reasoning, develops performance drop rate (PDR) to measure the relative decline in performance on question variations compared to the original questions (Li et al., 2024d). ErrorRadar uses error step accuracy and error category accuracy together to evaluate the multimodal error detection of MLLMs (Yan et al., 2024a).

Generative Evaluation, on the other hand, measures a M(LLM)'s ability to produce detailed explanations or solve problems from scratch. This evaluation type is gaining traction, particularly for complex mathematical tasks where step-by-step solutions are required. For instance, MathVerse, which modifies problems with varying degrees of information content in multi-modality, employs GPT-4 to score each key step in the reasoning process generated by MLLMs (Zhang et al., 2024f). CHAMP proposes a solution evaluation pipeline where GPT-4 is utilized as a grader for the answer summary, given the ground truth answer (Mao et al., 2024).

Due to page limit, more details of both types of evaluation metrics can be seen in Appendix D.

2.5 Training Data

The training of MLLMs for mathematical reasoning relies on a carefully orchestrated integration of *instruction design*, *data scale*, and *task diversity* to ensure robust and generalizable performance. Central to this process is the **design of instruction sets**, which are structured to bridge symbolic, textual, and visual reasoning (Toshniwal et al., 2024b,a). These instructions progressively escalate in complexity, starting from foundational arithmetic to advanced domains like calculus and linear algebra, ensuring models build skills incrementally. Each problem can be accompanied by explicit step-by-step explanations, enabling models to learn logical sequencing and self-correction (Zhang et al., 2024g; Tang et al., 2024b; Liang et al., 2024c).

The **scale of pre-training data** also plays an equally critical role. Models are exposed to terabytes of data sourced from textbooks, research papers (*e.g.,* arXiv), online educational platforms (*e.g.,* Khan Academy), and synthetically generated problems. A significant portion (10–30%) of the pretraining corpus is dedicated to mathematical content, with specialized datasets ensuring coverage of niche topics. While scaling to trillion-token corpora enhances robustness, rigorous filtering mechanisms, such as self-supervised quality checks, are applied to eliminate noise, including incorrect solutions or irrelevant content (Shao et al., 2024; Qwen, 2024; Yue et al., 2024c).

Finally, the **variety of mathematical tasks** ensures models adapt to diverse challenges. Training spans core domains like algebra and geometry, as well as cross-disciplinary applications (*e.g.,*

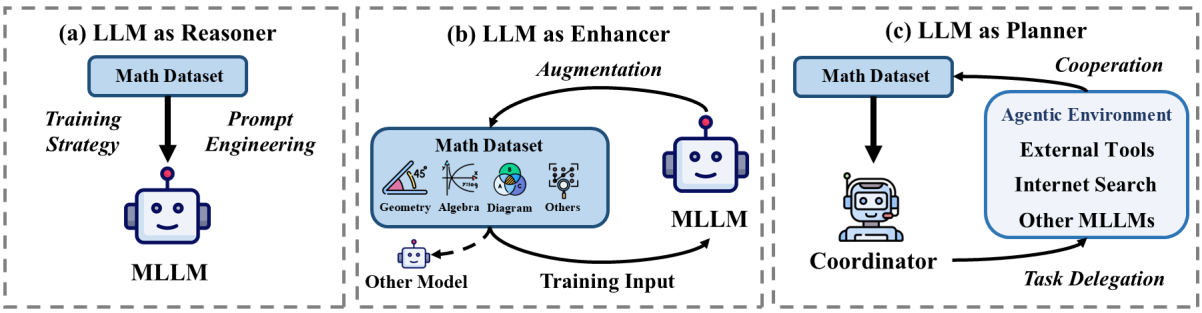

Figure 4: The illustration of the comparisons among three paradigms of (M)LLM-based mathematical reasoning.

physics-based calculus problems). Tasks are presented in multiple formats: closed-ended questions (*e.g.,* solving equations), open-ended prompts (*e.g.,* deriving proofs), and error-analysis exercises that require identifying and correcting flawed reasoning (Lu et al., 2022b; Yan et al., 2025).

For example, G-LLaVA (Gao et al., 2023) focuses on solving geometry problems by extracting visual features from geometric figures and jointly modeling them with text descriptions, allowing the model to understand key elements (*e.g.,* points, lines, angles) in geometric figures and their relationship with text descriptions. MAVIS (Zhang et al., 2024g) features an automatic data generation engine that can quickly generate large-scale, high-quality multimodal mathematical datasets, addressing the problem of data scarcity. It also uses instruction fine-tuning to teach the model how to decompose complex mathematical problems and generate reasonable reasoning steps (*esp.,* MAVIS-Instruct includes 834k visual math problems with CoT rationales). Math-LLaVA (Shi et al., 2024) uses the MathV360K multimodal dataset (360k instances), which covers multiple mathematical domains to gradually improve the model's mathematical reasoning ability through bootstrapping and further optimize the model using generated data.

3 Methodology Perspective

3.1 Overview & Findings

MLLMs have been leveraged in various ways to tackle the broad spectrum of mathematical reasoning tasks. Based on our comprehensive review of recent methodologies (summarized in Table 5 from Appendix E), we classify the works into three distinct paradigms: LLM as Reasoner (Sec.3.2), LLM as Enhancer (Sec.3.3), and LLM as Planner (Sec.3.4), and finally provide a in-depth comparison of technical distinctions (Sec.3.5).

Findings. First, single-modality settings dominate the current landscape of method-oriented research, with the majority focusing solely on algebraic tasks. However, since 2024, multimodal approaches have been increasingly incorporated, expanding the scope of mathematical reasoning to include geometry, diagrams, and even broader mathematical concepts. This shift signals a growing interest in enhancing model robustness through multimodal learning, which can address the diverse nature of mathematical problems. Second, regarding the evaluated tasks, problem-solving and proving are gaining prominence, while some research also focuses on error detection or others (*e.g.*, ReFAug includes error correction and follow-up QA as evaluation tasks (Zhang et al., 2024j)). Finally, in terms of the role of LLMs, Reasoner is the most common role, followed by Enhancer, while Planner remains less explored but holds promise due to recent advancements in multi-agent intelligence.

3.2 LLM as Reasoner

Definition. In the *Reasoner* paradigm, M(LLM)s harness their inherent reasoning capabilities to solve mathematical problems, as shown in Figure 4 (a). This can either involve fine-tuning existing LLMs on task-specific datasets or utilizing zero-shot or few-shot learning strategies. These models utilize advanced semantic understanding and reasoning techniques, such as symbolic manipulation, logical deduction, and multi-step reasoning.

Examples. Deng et al. (2023) develops a unified framework for answer calibration that integrates step-level and path-level strategies on multi-step reasoning of LLMs. MATH-SHEPHERD serves as a process-oriented math verifier, which assigns a reward score to each step of the LLM's outputs on math questions (Wang et al., 2024c). As for multimodal approaches, Math-PUMA introduces progressive upward multimodal alignment strat-

egy for reasoning-enhanced training (Zhuang et al., 2024); Math-LLaVA, a LLaVA-1.5-based model, directly bootstraps mathematical reasoning via fine-tuned on 360K high-quality math QA pairs, which can ensure the depth and breadth of multimodal mathematical problems (Shi et al., 2024); STIC develops a two-stage self-training pipeline (consisting of Image Comprehension Self-Training phase & Description-Infused Fine-Tuning phase) for enhancing visual comprehension (Deng et al., 2024); VCAR emphasizes on the visual-centric supervision, thus proposing a similar two-step training pipleine which handles the visual description generation task first, followed by mathematical rationale generation task (Jia et al., 2024).

Summary & Outlook. This paradigm has shown significant promise, particularly in solving problems requiring multiple steps of reasoning. However, despite improvements, issues with robustness remain, particularly with zero-shot reasoning tasks. Future work should focus on combining reasoning with structured knowledge retrieval systems and enhancing models' ability to reason effectively across diverse domains, especially in multimodal contexts (Fan et al., 2024b; Pan et al., 2023).

3.3 LLM as Enhancer

Definition. In the *Enhancer* paradigm, M(LLM)s are primarily used to augment data, thereby enabling improvements in mathematical reasoning, as illustrated in Figure 4 (b). This can be achieved by synthesizing new training data, refining existing datasets, or introducing new variations that target specific problem-solving abilities (Li et al., 2022). Data augmentation can include paraphrasing mathematical problems, adding noise to mathematical expressions, or generating problem variants for underrepresented cases.

Examples. A typical example of a single-modality enhancement approach is Masked Thought, which introduces perturbations to the input and randomly masks tokens within the chain of thought during training (Chen et al., 2024a). MathGenie, which aims to generate diverse and reliable math problems and solution from a small-scale dataset, leverages a solution augmentation model to iteratively create new solutions from existing ones (Lu et al., 2024b). For multimodal methods, AlphaGeometry proves most olympiad-level mathematical theorems, via trained from scratch on large-scale synthetic data guiding the symbolic deuction (Trinh et al., 2024); LogicSolver introduces interpretable formula-based tree-structure for each solution equation (Yang et al., 2022); InfiMM-Math achieves the exceptional performance as it is trained on a large-scale multimodal interleaved math dataset developed and validated by LLMs such as LLaMA3-70B-Instruct (Han et al., 2024); DFE-GPS constructs its synthetic training set, which integrates visual features and geometric formal language (Zhang et al., 2024i).

Summary & Outlook. This paradigm offers substantial performance improvements by enriching the training set. However, challenges remain in ensuring the diversity and relevance of the generated data. Moreover, while text-based augmentation methods have proven effective, the potential for multimodal augmentation is still underexplored. Future research should focus on advancing multimodal data augmentation techniques, especially for tasks that require interaction between visual and textual modalities (Xiao et al., 2023).

3.4 LLM as Planner

Definition. In the *Planner* paradigm, M(LLM)s are treated as coordinators that guide the solution of complex mathematical problems by delegating tasks to other models or tools, as illustrated in Figure 4 (c). This includes scenarios where multiple agents or models collaborate to achieve a single objective, thereby enhancing the performance of mathematical problem-solving through cooperative interactions. These models often work in environments with multiple steps or require iterative refinement of solutions.

Examples. A notable tool-integrated agent is ToRA, which plans the sequential use of natural language rationale and program-based tools synergistically to solve mathematical problems in an optimal manner (Gou et al., 2023). Additionally, COPRA simulates a single agent-like reasoning mechanism where GPT-4 proposes tactic applications within a stateful backtracking search, leveraging feedback from the proof environment (Thakur et al., 2024). This can also extend to multimodal scenarios, as seen in Chameleon, which serves as an AI system that augments MLLMs with plug-and-play modules for compositional reasoning, leveraging an LLM-based planner to assemble tools for complex tasks (Lu et al., 2024a). Furthermore, Visual Sketchpad presents the concept of sketching as a ubiquitous tool used by humans for communication, ideation, and problem-solving. Hence, MLLMs can enable external tools (*e.g.*, matplotlib)

Aspect	LLM as Reasoner	LLM as Enhancer	LLM as Planner
Data Interaction Patterns			
Input-Output Relation	End-to-end mapping (Problem → Answer)	Data augmentation pipeline (Raw data → Enhanced data)	Dynamic workflow planning (Problem → Plan → Subtasks)
External Dependencies	Low (Self-contained reasoning)	Medium (Data distribution dependent)	High (Requires toolchain integration)
Pros & Cons			
Advantages	Transparent reasoning & Strong interpretability	Improves generalization & Handles data scarcity	Breaks capability boundaries & Enables complex task solving
Limitations	Error-prone in complex reasoning	May introduce semantic biases	High system complexity & Increased latency

Table 2: Comparisons among the three methodology paradigms.

to generate intermediate sketches to aid in reasoning, which includes an iterative interaction process with an environment (Hu et al., 2024). Although there has been much work on Compositional Visual Reasoning in the past (Gupta and Kembhavi, 2023; Surís et al., 2023; Yao et al., 2022), Visual Sketchpad is the first work that integrates the planning capabilities of MLLMs with the real gap of mathematical reasoning settings (*i.e.*, sketch-based reasoning involving visuo-spaital concepts).

Summary & Outlook. While the Planner paradigm introduces significant improvements, particularly for complex tasks that require multi-agent collaboration, it remains a relatively under-explored area (Xi et al., 2023; Guo et al., 2024b). There is potential for further improvement in task decomposition, agent cooperation strategies, and integration of diverse computational tools. Future work will likely focus on refining these planning strategies, especially for multimodal systems that can jointly leverage visual and textual knowledge to solve more intricate problems (Xie et al., 2024; Durante et al., 2024; Li et al., 2023b).

3.5 Paradigm Comparison

As summarized in Table 2, we list the differences between the three paradigms to provide the community with a more comprehensive understanding of the latest technical distinctions. These three paradigms show a progressive development logic: Reasoner focuses on intrinsic model capabilities, Enhancer targets data optimization, and Planner moves towards system-level intelligent collaboration. In practice, we also anticipate adopting a hybrid approach (*e.g.,* using Enhancer to generate augmented data to train Reasoner, then coordinating multiple Reasoner modules via Planner to solve complex problems). This layered architecture may become the core design paradigm for future multimodal mathematical reasoning systems.

4 Challenges

In the realm of MLLMs for mathematical reasoning, the following key challenges persist that hinder their full potential. Addressing these challenges is essential for advancing MLLMs toward more robust and flexible systems that can better support mathematical reasoning in real-world settings.

❶ **Lack of High-Quality, Diverse, and Large-Scale Multimodal Datasets**. As discussed in Section 2.5, current multimodal mathematical reasoning datasets face tripartite limitations in quality (*e.g.,* misaligned text-image pairs), scale (insufficient advanced topic coverage), and task diversity (overemphasis on problem-solving versus error diagnosis or theorem proving). For instance, most datasets focus on question answering but lack annotations for error tracing steps or formal proof generation, while synthetic datasets often exhibit domain bias (Wang et al., 2024a; Lu et al., 2023). Three concrete solutions emerge: i) Develop hybrid dataset construction pipelines combining expert-curated problems with AI-augmented task variations; ii) Implement cross-task knowledge distillation, where models trained on proof generation guide error diagnosis through attention pattern transfer; iii) Leverage automated frameworks quality-controlled multimodal expansion to systematically generate diverse task formats (*e.g.,* converting proof exercises into visual dialogues). More discussion on data bottlenecks in Appendix F.1.

❷ **Insufficient Visual Reasoning**. Many math problems require extracting and reasoning over visual content, such as charts, tables, or geometric diagrams. Current models struggle with intricate visual details, such as interpreting three-dimensional geometry or analyzing irregularly structured tables (Zhang et al., 2024f). Hence, it may be beneficial to introduce enhanced visual feature extraction modules and integrate scene graph representations for better reasoning over complex visual elements (Ibrahim et al., 2024; Guo et al., 2024c).

❸ **Reasoning Beyond Text and Vision**. While the current research focus on the combination of text and vision, mathematical reasoning in real-world applications often extends beyond these two modalities. For instance, audio explanations, interactive problem-solving environments, or dynamic simulations might play a role in some tasks. Current models are not well-equipped to handle such diverse inputs (Abrahamson et al., 2020; Jusslin et al., 2022). To address this, datasets should be expanded to include more diverse modalities, such as audio, video, and interactive tools. MLLMs should also be designed with flexible architectures capable of processing and reasoning over multiple types of inputs, allowing for a richer representation of mathematical problems (Dasgupta et al., 2023).

❹ **Limited Domain Generalization**. Mathematical reasoning spans many domains, such as algebra, geometry, diagram and commonsense, each with its own specific requirements for problem-solving (Liu et al., 2023b; Lu et al., 2022b). Math-LLMs that perform well in one domain often fail to generalize across others, which can limit their utility. By pretraining and fine-tuning Math-LLMs on a wide array of problem types, models may handle cross-domain tasks more effectively, improving their ability to generalize across different mathematical topics and problem-solving strategies. We extend more discussion on limited domain generalization in multimodal contexts in Appendix F.2.

❺ **Error Feedback Limitations**. Mathematical reasoning involves various types of errors, such as calculation mistakes, logical inconsistencies, and misinterpretations of the problem. Currently, MLLMs lack mechanisms to detect, categorize, and correct these errors effectively, which can result in compounding mistakes throughout the reasoning process (Yan et al., 2024a; Li et al., 2024e). A potential solution is to integrate error detection and classification modules that can identify errors at each step of the reasoning process. Besides, multi-agent collaboration mechanism could be introduced, via involving multiple agents collaborating by exchanging feedback and collectively refining the reasoning process (Xu et al., 2024d). We extend more discussion on error feedback limitation in multimodal contexts in Appendix F.3.

❻ **Integration with Real-world Educational Needs**. Existing benchmarks and models often overlook real-world educational contexts, such as how students use draft work, like handwritten notes or diagrams, to solve problems (Xu et al., 2024c; Wang et al., 2024d). These real-world elements are crucial for understanding how humans approach mathematical reasoning (Mouchere et al., 2011; Gervais et al., 2024). By incorporating draft notes, handwritten calculations, and dynamic problem-solving workflows into the training data, MLLMs can be tailored to provide more accurate and contextually relevant feedback for students.

❼ **Test-Time Scaling Technique in Multimodal Context**. While foundation models increasingly adopt test-time scaling techniques (*e.g.,* dynamic architecture adaptation), their integration with multimodal mathematical reasoning remains underexplored and suboptimal. For example, current implementations like o1 (Jaech et al., 2024) or DeepSeek-R1 (Guo et al., 2025) struggle to dynamically allocate computational resources based on math problem complexity across modalities, such as deciding when to prioritize symbolic computation over visual parsing for optimization problems. Future work should focus on two directions: i) Develop modality-aware scaling controllers that jointly consider problem type, visual complexity, and required mathematical operations to optimize dynamic architecture decisions; ii) Create lightweight meta-optimization layers that can adjust model capacity allocation (*e.g.,* expert selection in MoE systems) through real-time analysis of multimodal problem-solving workflows (Xu et al., 2025a; Besta et al., 2025). Such advancements could enable more efficient trade-offs between accuracy and computational cost in deployed systems. We also discuss how test-time scaling techniques can tackle the other challenges in Appendix F.4.

5 Conclusion

In this survey, we have provided a comprehensive overview of the progress and challenges in mathematical reasoning within the context of MLLMs. We highlighted the significant advances in the development of Math-LLMs and the growing importance of multimodal integration for solving complex reasoning tasks. We identified five key challenges that are crucial for the continued development of AGI systems capable of performing sophisticated mathematical reasoning tasks. As research continues to advance, it is essential to focus on these challenges to unlock the full potential of LLMs in multimodal settings. We hope this survey provides insights to guide future LLM research, ultimately leading to more effective and human-like mathematical reasoning capabilities in AI systems.

Limitations

Despite our best efforts to ensure comprehensive coverage of the published works, it is possible that some relevant studies were overlooked. Additionally, human errors could have occurred during the categorization or referencing of papers in the survey. To minimize such errors, we made a concerted effort to gather studies from multiple sources and performed a multiple-round checking process. While minor inconsistencies or omissions may still exist, we believe this survey represents the most comprehensive review of MLLM-based mathematical reasoning to date, effectively capturing key research trends and highlighting ongoing challenges.

References

Dor Abrahamson, Mitchell J Nathan, Caro Williams-Pierce, Candace Walkington, Erin R Ottmar, Hortensia Soto, and Martha W Alibali. 2020. The future of embodied design for mathematics teaching and learning. In Frontiers in Education, volume 5, page 147. Frontiers Media SA.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. arXiv preprint arXiv:2402.00157.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:1905.13319.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. arXiv preprint arXiv:2310.10631.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. 2024. A survey of multimodal large language model from a data-centric perspective. arXiv preprint arXiv:2405.16640.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, et al. 2025. Reasoning language models: A blueprint. arXiv preprint arXiv:2501.11223.

Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. 2024a. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. arXiv preprint arXiv:2403.02178.

Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. 2025. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. arXiv preprint arXiv:2502.07154.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. arXiv preprint arXiv:2212.02746.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. arXiv preprint arXiv:2105.14517.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv preprint arXiv:2405.16473.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7889–7901.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271.

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Joanna Matthiesen, Kevin Smith, and Joshua B Tenenbaum. 2024. Evaluating large vision-and-language models on children's mathematical olympiads. arXiv preprint arXiv:2406.15736.

Ethan Chern, Haoyang Zou, Xuefeng Li, Jiewen Hu, Kehua Feng, Junlong Li, and Pengfei Liu. 2023. Generative ai for math: Abel. https://github.com/GAIR-NLP/abel.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. 2024. U-math: A university-level benchmark for evaluating mathematical skills in llms. arXiv preprint arXiv:2412.03205.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Adrian Cosma, Ana-Maria Bucur, and Emilian Radoi. 2024. Romath: A mathematical reasoning benchmark in romanian. arXiv preprint arXiv:2409.11074.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092.

Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. 2023. Collaborating with language models for embodied reasoning. arXiv preprint arXiv:2302.00763.

Arash Gholami Davoodi, Seyed Pouyan Mousavi Davoudi, and Pouya Pezeshkpour. 2024. Llms are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of llms. arXiv preprint arXiv:2406.05194.

Shumin Deng, Ningyu Zhang, Nay Oo, and Bryan Hooi. 2023. Towards a unified view of answer calibration for multi-step reasoning. arXiv preprint arXiv:2311.09101.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. arXiv preprint arXiv:2405.19716.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. arXiv preprint arXiv:2405.12205.

Prakhar Dixit and Tim Oates. 2024. Sbi-rag: Enhancing math word problem solving for students through schema-based instruction and retrieval-augmented generation. arXiv preprint arXiv:2410.13293.

Duolingo. 2024. Duolingo official platfrom.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. arXiv preprint arXiv:2401.03568.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. 2024a. Hardmath: A benchmark dataset for challenging problems in applied mathematics. arXiv preprint arXiv:2410.09988.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024b. A survey on rag meeting llms: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 6491–6501.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. arXiv preprint arXiv:2406.18321.

Shengyu Feng, Xiang Kong, Shuang Ma, Aonan Zhang, Dong Yin, Chong Wang, Ruoming Pang, and Yiming Yang. 2024. Step-by-step reasoning for math problems via twisted sequential monte carlo. arXiv preprint arXiv:2410.01920.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024a. Isobench: Benchmarking multimodal foundation models on isomorphic representations. arXiv preprint arXiv:2404.01266.

Jiayi Fu, Lei Lin, Xiaoyang Gao, Pengli Liu, Zhengzong Chen, Zhirui Yang, Shengnan Zhang, Xue Zheng, Yan Li, Yuliang Liu, et al. 2023. Kwaiyiimath: Technical report. arXiv preprint arXiv:2310.07488.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pages 148–166. Springer.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. arXiv preprint arXiv:2410.07985.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370.

Philippe Gervais, Asya Fadeeva, and Andrii Maksai. 2024. Mathwriting: A dataset for handwritten mathematical expression recognition. arXiv preprint arXiv:2404.10690.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. arXiv preprint arXiv:2309.17452.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning. In The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Siyuan Guo, Aniket Didolkar, Nan Rosemary Ke, Anirudh Goyal, Ferenc Huszár, and Bernhard Schölkopf. 2024a. Learning beyond pattern matching? assaying mathematical understanding in llms. arXiv preprint arXiv:2405.15485.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024c. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph. Complex & Intelligent Systems, 10(5):7063–7076.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A challenging multi-modal mathematical reasoning benchmark. arXiv preprint arXiv:2410.14702.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14953–14962.

Vernon Toh Yan Han, Ratish Puduppully, and Nancy F Chen. 2023. Veritymath: Advancing mathematical reasoning by self-verification through unit consistency. arXiv preprint arXiv:2311.07172.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, et al. 2024. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. arXiv preprint arXiv:2409.12568.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. arXiv preprint arXiv:2402.14008.

Stephen J Hegedus and David O Tall. 2015. Foundations for the future: The potential of multimodal technologies for learning mathematics. In Handbook of international research in mathematics education, pages 543–562. Routledge.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403.

Litian Huang, Xinguo Yu, Feng Xiong, Bin He, Shengbing Tang, and Jiawen Fu. 2024a. Hologram reasoning for solving algebra problems with geometry diagrams. arXiv preprint arXiv:2408.10592.

Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, and Benyou Wang. 2024b. Mamo: a mathematical modeling benchmark with solvers. arXiv preprint arXiv:2405.13144.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2024c. Key-point-driven data synthesis with its enhancement on mathematical reasoning. arXiv preprint arXiv:2403.02333.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. 2024d. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. arXiv preprint arXiv:2409.09039.

Nourhan Ibrahim, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. 2024. A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges. Discover Artificial Intelligence, 4(1):76.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. arXiv preprint arXiv:2404.14604.

Hyoungwook Jin, Yoonsu Kim, Yeon Su Park, Bekzat Tilekbay, Jinho Son, and Juho Kim. 2024. Using large language models to diagnose math problem-solving skills at scale. In Proceedings of the Eleventh ACM Conference on Learning@ Scale, pages 471–475.

Bert Jonsson, Julia Mossegård, Johan Lithner, and Linnea Karlsson Wirebring. 2022. Creative mathematical reasoning: Does need for cognition matter? Frontiers in Psychology, 12:797807.

Sofia Jusslin, Kaisa Korpinen, Niina Lilja, Rose Martin, Johanna Lehtinen-Schnabel, and Eeva Anttila. 2022. Embodied learning and teaching approaches in language education: A mixed studies review. Educational Research Review, 37:100480.

Jikun Kang, Xin Zhe Li, Xi Chen, Amirreza Kazemi, Qianyi Sun, Boxing Chen, Dong Li, Xu He, Quan He, Feng Wen, et al. 2024. Mindstar: Enhancing math reasoning in pre-trained llms at inference time. arXiv preprint arXiv:2405.16265.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. arXiv preprint arXiv:2312.12241.

KhanAcademy. 2024. Khanmigo official platfrom.

JB Kim, Hazel Kim, Joonghyuk Hahn, and Yo-Sub Han. 2023. Athena: Mathematical reasoning with thought expansion. arXiv preprint arXiv:2311.01036.

Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-lm: A dynamic benchmark for mathematical reasoning on large language models. arXiv preprint arXiv:2406.12572.

Guillaume Lample, Timothee Lacroix, Marie-Anne Lachaux, Aurelien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. 2022. Hypertree proof search for neural theorem proving. Advances in neural information processing systems, 35:26337–26349.

Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, and Kang Min Yoo. 2024. Token-supervised value models for enhancing mathematical reasoning capabilities of large language models. arXiv preprint arXiv:2407.12863.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2024. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. arXiv preprint arXiv:2404.04735.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems, 35:3843–3857.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. Ai Open, 3:71–90.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024a. Common 7b language models already possess strong math capabilities. arXiv preprint arXiv:2403.04706.

Chengpeng Li, Guanting Dong, Mingfeng Xue, Ru Peng, Xiang Wang, and Dayiheng Liu. 2024b. Dotamath: Decomposition of thought with code assistance and self-correction for mathematical reasoning. arXiv preprint arXiv:2407.04078.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2024c. Mugglemath: Assessing the impact of query and response augmentation on math reasoning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10230–10258.

Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023a. Adapting large language models for education: Foundational capabilities, potentials, and challenges. arXiv preprint arXiv:2401.08664.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024d. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. arXiv preprint arXiv:2402.19255.

Wenhua Li, Tao Zhang, Rui Wang, Shengjun Huang, and Jing Liang. 2023b. Multimodal multi-objective optimization: Comparative study of the state-of-the-art. Swarm and Evolutionary Computation, 77:101253.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. arXiv preprint arXiv:2501.05366.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024e. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. arXiv preprint arXiv:2406.00755.

Zenan Li, Zhi Zhou, Yuan Yao, Yu-Feng Li, Chun Cao, Fan Yang, Xian Zhang, and Xiaoxing Ma. 2024f. Neuro-symbolic data generation for math reasoning. arXiv preprint arXiv:2412.04857.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024g. A survey on deep learning for theorem proving. arXiv preprint arXiv:2404.09939.

Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024h. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. arXiv preprint arXiv:2408.11397.

Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Fan-Hu Zeng, Jian Xu, Jia-Xin Zhang, and Cheng-Lin Liu. 2024i. Cmmath: A chinese multi-modal math skill evaluation benchmark for foundation models. arXiv preprint arXiv:2407.12023.

Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023c. Lans: A layout-aware neural solver for plane geometry problem. arXiv preprint arXiv:2311.16476.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2024a. Quantifying & modeling multimodal interactions: An information decomposition framework. Advances in Neural Information Processing Systems, 36.

Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. 2024b. Improving llm reasoning through scaling inference computation with collaborative verification. arXiv preprint arXiv:2410.05318.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023a. Unimath: A foundational and multimodal mathematical reasoner. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7126–7133.

Zhenwen Liang, Dian Yu, Xiaoman Pan, Wenlin Yao, Qingkai Zeng, Xiangliang Zhang, and Dong Yu. 2023b. Mint: Boosting generalization in mathematical reasoning via multi-view fine-tuning. arXiv preprint arXiv:2307.07951.

Zhenwen Liang, Dian Yu, Wenhao Yu, Wenlin Yao, Zhihan Zhang, Xiangliang Zhang, and Dong Yu. 2024c. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. arXiv preprint arXiv:2405.19444.

Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2021. Mwp-bert: Numeracy-augmented pre-training for math word problem solving. arXiv preprint arXiv:2107.13435.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024. Rho-1: Not all tokens are what you need. arXiv preprint arXiv:2404.07965.

Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew Chi-Chih Yao. 2024a. Augmenting math word problems via iterative question composing. arXiv preprint arXiv:2401.09003.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024b. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. arXiv preprint arXiv:2405.12209.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023a. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956.

Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, et al. 2023b. Mathematical language models: A survey. arXiv preprint arXiv:2312.07622.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024c. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. arXiv preprint arXiv:2409.02834.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024d. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. arXiv preprint arXiv:2402.17644.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. arXiv preprint arXiv:2105.04165.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024a. Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems, 36.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022a. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2022b. A survey of deep learning for mathematical reasoning. arXiv preprint arXiv:2212.10535.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024b. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. arXiv preprint arXiv:2402.16352.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024c. Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code. Preprint, arXiv:2410.08196.

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

Jingkun Ma, Runzhe Zhan, Derek F Wong, Yang Li, Di Sun, Hou Pong Chan, and Lidia S Chao. 2024. Visaidmath: Benchmarking visual-aided mathematical reasoning. arXiv preprint arXiv:2410.22995.

Yujun Mao, Yoon Kim, and Yilun Zhou. 2024. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. arXiv preprint arXiv:2401.06961.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022. Lila: A unified benchmark for mathematical reasoning. arXiv preprint arXiv:2210.17517.

MistralAI. 2024. Mathstral official platform.

MoonshotAI. 2024. k0-math official platform.

Harold Mouchere, Christian Viard-Gaudin, Dae Hwan Kim, Jin Hyung Kim, and Utpal Garain. 2011. Crohme2011: Competition on recognition of online handwritten mathematical expressions. In 2011 international conference on document analysis and recognition, pages 1497–1500. IEEE.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In Informatics, volume 11, page 57. MDPI.

Bolin Ni, JingCheng Hu, Yixuan Wei, Houwen Peng, Zheng Zhang, Gaofeng Meng, and Han Hu. 2024. Xwin-lm: Strong and scalable alignment practice for llms. arXiv preprint arXiv:2405.20335.

OpenAI. 2024. Gpt-4o system card.

Kay L O'Halloran. 2015. The language of learning mathematics: A multimodal perspective. The Journal of Mathematical Behavior, 40:63–74.

Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. arXiv preprint arXiv:2308.06374.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. arXiv preprint arXiv:2409.00147.

Gabriel Poesia, David Broman, Nick Haber, and Noah D Goodman. 2024. Learning formal mathematics from intrinsic motivation. arXiv preprint arXiv:2407.00695.

Stanislas Polu and Ilya Sutskever. 2021. Generative language modeling for automated theorem proving. arXiv preprint arXiv:2009.03393.

Jinghui Qin, Zhicheng Yang, Jiaqi Chen, Xiaodan Liang, and Liang Lin. 2023. Template-based contrastive distillation pretraining for math word problem solving. IEEE Transactions on Neural Networks and Learning Systems.

Qwen. 2024. Qwen2-math technical report.

AM Rahman, Junyi Ye, Wei Yao, Wenpeng Yin, and Guiling Wang. 2024. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems. arXiv preprint arXiv:2410.18921.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can llms master math? investigating large language models on math stack exchange. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2316–2320.

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

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. arXiv preprint arXiv:2406.17294.

Shiven Sinha, Ameya Prabhu, Ponnurangam Kumaraguru, Siddharth Bhat, and Matthias Bethge. 2024. Wu's method can boost symbolic ai to rival silver medalists and alphageometry to outperform gold medalists at imo geometry. arXiv preprint arXiv:2404.06405.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. arXiv preprint arXiv:2311.07594.

SquirrelAiLearning. 2024. Squirrel ai official platfrom.

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. Evaluating llms' mathematical reasoning in financial document question answering. In Findings of the Association for Computational Linguistics ACL 2024, pages 3853–3878.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024a. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained

classification. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1358–1375.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024b. Scieval: A multi-level large language model evaluation benchmark for scientific research. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19053–19061.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Conghui He, Zenan Zhou, and Wentao Zhang. 2024c. Beats: Optimizing llm mathematical capabilities with backverify and adaptive disambiguate based efficient tree search. arXiv preprint arXiv:2409.17972.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11888–11898.

TALEducation. 2023. Mathgpt official platform.

Jiamin Tang, Chao Zhang, Xudong Zhu, and Mengchi Liu. 2024a. Tangram: A challenging benchmark for geometric element recognizing. arXiv preprint arXiv:2408.13854.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024b. Mathscale: Scaling instruction tuning for mathematical reasoning. arXiv preprint arXiv:2403.02884.

Amitayush Thakur, George Tsoukalas, Yeming Wen, Jimmy Xin, and Swarat Chaudhuri. 2024. An in-context learning agent for formal theorem-proving. In First Conference on Language Modeling.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. arXiv preprint arXiv:2407.13690.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024a. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. arXiv preprint arXiv:2410.01560.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024b. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. arXiv preprint arXiv:2402.10176.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. Nature, 625(7995):476–482.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. In AI for Math Workshop@ ICML 2024.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. arXiv preprint arXiv:2402.14804.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023a. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. arXiv preprint arXiv:2310.03731.

Lei Wang, Shan Dong, Yuhui Xu, Hanze Dong, Yalu Wang, Amrita Saha, Ee-Peng Lim, Caiming Xiong, and Doyen Sahoo. 2024b. Mathhay: An automated benchmark for long-context mathematical reasoning in llms. arXiv preprint arXiv:2410.04698.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9426–9439.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024d. Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023b. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024e. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. arXiv preprint arXiv:2401.06805.

Yixu Wang, Wenpin Qian, Hong Zhou, Jianfeng Chen, and Kai Tan. 2023c. Exploring new frontiers of deep learning in legal practice: A case study of large language models. International Journal of Computer Science and Information Technology, 1(1):131–138.

Chenrui Wei, Mengzhou Sun, and Wei Wang. 2024. Proving olympiad algebraic inequalities without human demonstrations. arXiv preprint arXiv:2406.14219.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023a. Multimodal large language models: A survey. In 2023 IEEE International Conference on Big Data (BigData), pages 2247–2256. IEEE.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Ting Wu, Xuefeng Li, and Pengfei Liu. 2024a. Progress or regress? self-improvement reversal in post-training. arXiv preprint arXiv:2407.05013.

Zhenyu Wu, Meng Jiang, and Chao Shen. 2024b. Get an a in math: Progressive rectification prompting. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19288–19296.

Zijian Wu, Suozhi Huang, Zhejian Zhou, Huaiyuan Ying, Jiayu Wang, Dahua Lin, and Kai Chen. 2024c. Internlm2. 5-stepprover: Advancing automated theorem proving via expert iteration on large-scale lean problems. arXiv preprint arXiv:2410.15700.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. 2023. Multimodal data augmentation for image captioning using diffusion models. In Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications, pages 23–33.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. arXiv preprint arXiv:2402.15116.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024a. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. arXiv preprint arXiv:2405.14333.

Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. 2024b. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. arXiv preprint arXiv:2408.08152.

Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, et al. 2024. Building math agents with multi-turn iterative preference learning. arXiv preprint arXiv:2409.02392.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025a. Towards large reasoning models: A survey of reinforced reasoning with large language models. arXiv preprint arXiv:2501.09686.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. 2024a. Large language models for education: A survey. arXiv preprint arXiv:2405.13001.

Liang Xu, Hang Xue, Lei Zhu, and Kangkang Zhao. 2024b. Superclue-math6: Graded multi-step math reasoning benchmark for llms in chinese. arXiv preprint arXiv:2401.11819.

Tianlong Xu, Richard Tong, Jing Liang, Xing Fan, Haoyang Li, and Qingsong Wen. 2024c. Foundation models for education: Promises and prospects. arXiv preprint arXiv:2405.10959.

Tianlong Xu, Yi-Fan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. 2024d. Ai-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. arXiv preprint arXiv:2409.09403.

Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. 2024e. Can llms solve longer math word problems better? arXiv preprint arXiv:2405.14804.

Xin Xu, Jiaxin Zhang, Tianhao Chen, Zitong Chao, Jishan Hu, and Can Yang. 2025b. Ugmathbench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models. arXiv preprint arXiv:2501.13766.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, et al. 2024f. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. arXiv preprint arXiv:2404.02893.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pages 4163–4167.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. 2024a. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. arXiv preprint arXiv:2410.04509.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal large language models can significantly advance scientific reasoning. arXiv preprint arXiv:2502.02871.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024b. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In Proceedings of the ACM on Web Conference 2024, pages 4006–4017.

Yuchen Yan, Jin Jiang, Yang Liu, Yixin Cao, Xin Xu, Xunliang Cai, Jian Shao, et al. 2024c. S3c-math: Spontaneous step-level self-correction makes large language models better mathematical reasoners. arXiv preprint arXiv:2409.01524.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. arXiv preprint arXiv:2306.06031.

Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, Yuxiao Dong, and Jie Tang. 2024b. Mathglm-vision: Solving mathematical problems with multi-modal large language model. arXiv preprint arXiv:2409.13729.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. 2023b. Gpt can solve mathematical problems without a calculator. arXiv preprint arXiv:2309.03241.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. 2022. Logicsolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. arXiv preprint arXiv:2205.08232.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. National Science Review, page nwae403.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. arXiv preprint arXiv:2402.06332.

Dian Yu, Baolin Peng, Ye Tian, Linfeng Song, Haitao Mi, and Dong Yu. 2024a. Siam: Self-improving code-assisted mathematical reasoning of large language models. arXiv preprint arXiv:2408.15565.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024b. Natural language reasoning, a survey. ACM Computing Surveys, 56(12):1–39.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of CVPR.

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2409.02813.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024c. Mammoth2: Scaling instructions from the web. arXiv preprint arXiv:2405.03548.

Liang Zeng, Liangjun Zhong, Liang Zhao, Tianwen Wei, Liu Yang, Jujie He, Cheng Cheng, Rui Hu, Yang Liu, Shuicheng Yan, et al. 2024. Skywork-math: Data scaling laws for mathematical reasoning in large language models–the story goes on. arXiv preprint arXiv:2407.08348.

Beichen Zhang, Kun Zhou, Xilin Wei, Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2024a. Evaluating and improving tool-augmented computation-intensive math reasoning. Advances in Neural Information Processing Systems, 36.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. 2024b. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. arXiv preprint arXiv:2410.02884.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024c. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601.

Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024d. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. arXiv preprint arXiv:2402.10104.

Mengxue Zhang, Zichao Wang, Zhichao Yang, Weiqi Feng, and Andrew Lan. 2023. Interpretable math word problem solution generation via step-by-step planning. arXiv preprint arXiv:2306.00784.

Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, Liang Lin, and Cheng-Lin Liu. 2024e. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. arXiv preprint arXiv:2407.07327.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024f. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In European Conference on Computer Vision, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. 2024g. Mavis: Mathematical visual instruction tuning with an automatic data engine. arXiv preprint arXiv:2407.08739.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In Proceedings of the 32nd ACM international conference on information and knowledge management, pages 4435–4439.

Yiming Zhang, Baoyi He, Shengyu Zhang, Yuhao Fu, Qi Zhou, Zhijie Sang, Zijin Hong, Kejing Yang, Wenjun Wang, Jianbo Yuan, et al. 2024h. Unconstrained model merging for enhanced llm reasoning. arXiv preprint arXiv:2410.13699.

Zeren Zhang, Jo-Ku Cheng, Jingyang Deng, Lu Tian, Jinwen Ma, Ziran Qin, Xiaokai Zhang, Na Zhu, and Tuo Leng. 2024i. Diagram formalization enhanced multi-modal geometry problem solver. arXiv preprint arXiv:2409.04214.

Zhihan Zhang, Tao Ge, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, and Meng Jiang. 2024j. Learn beyond the answer: Training language models with reflection for mathematical reasoning. arXiv preprint arXiv:2406.12050.

Wayne Xin Zhao, Kun Zhou, Zheng Gong, Beichen Zhang, Yuanhang Zhou, Jing Sha, Zhigang Chen, Shijin Wang, Cong Liu, and Ji-Rong Wen. 2022. Jiuzhang: A chinese pre-trained language model for mathematical problem understanding. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 4571–4581.

Xin Zhao, Kun Zhou, Beichen Zhang, Zheng Gong, Zhipeng Chen, Yuanhang Zhou, Ji-Rong Wen, Jing Sha, Shijin Wang, Cong Liu, et al. 2023. Jiuzhang 2.0: A unified chinese pre-trained language model for multi-task mathematical problem solving. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 5660–5672.

Xueliang Zhao, Xinting Huang, Wei Bi, and Lingpeng Kong. 2024. Sego: Sequential subgoal optimization for mathematical problem-solving. arXiv preprint arXiv:2310.12960.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, et al. 2024. Evaluation of openai o1: Opportunities and challenges of agi. arXiv preprint arXiv:2409.18486.

Kun Zhou, Beichen Zhang, Jiapeng Wang, Zhipeng Chen, Wayne Xin Zhao, Jing Sha, Zhichao Sheng, Shijin Wang, and Ji-Rong Wen. 2024a. Jiuzhang3.0: Efficiently improving mathematical reasoning by training small data synthesis models. arXiv preprint arXiv:2405.14365.

Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. 2024b. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. arXiv preprint arXiv:2408.07543.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024c. Lawgpt: A chinese legal knowledge-enhanced large language model. arXiv preprint arXiv:2406.04614.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024d. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. arXiv preprint arXiv:2407.08733.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2024e. Mathattack: Attacking large language models towards math solving ability. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19750–19758.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2022. Solving math word problems via cooperative reasoning induced language models. arXiv preprint arXiv:2210.16257.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2024. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. arXiv preprint arXiv:2408.08640.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. arXiv preprint arXiv:2411.00836.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. Information Fusion, 113:102606.

A Details of Math-LLMs' Progress

The rapid development of general-purpose LLMs has made significant advancements in natural language processing tasks. However, the development of domain-specific models remains a core requirement, as they are better equipped to handle specialized tasks that general models may not address effectively. This is particularly true in fields such as healthcare (Liu et al., 2023a; Nazi and Peng, 2024), law (Cui et al., 2023; Zhou et al., 2024c; Wang et al., 2023c), finance (Wu et al., 2023b; Yang et al., 2023a; Zhang and Yang, 2023), and urban science (Yan et al., 2024b; Zou et al., 2025; Yan and Lee, 2024), where domain-specific knowledge is critical for high accuracy and performance.

In the case of mathematical reasoning, general models may struggle with tasks that require a deep understanding of complex mathematical concepts, structures, and problem-solving steps (Yan et al., 2025). Therefore, the development of math-specific LLMs is of paramount importance, as these models are designed to enhance performance in mathematical reasoning, theorem proving, equation solving, and other math-intensive tasks.

Therefore, Table 4 provides a detailed overview of various math-specific LLMs (*i.e.*, Math-(LLMs), sorted by their release date. It includes information about the organization behind each model, the release date, publication details, language(s) supported, parameter size, evaluation benchmarks, and whether the model is open source.

Key findings are summarized as follows:

1. **Release Trends:** The models started emerging in 2020, with a significant increase in the number of releases from 2022 onward, indicating a growing interest in developing math-specific LLMs.

2. **Parameter Sizes:** There is a noticeable trend towards larger parameter sizes, with some models offering up to 130B parameters, reflecting the increasing computational capacity for handling complex mathematical tasks.

3. **Evaluation Benchmarks:** Many models are evaluated on popular benchmarks like GSM8K, MATH, and MMLU, highlighting the focus on improving performance across well-established mathematical reasoning datasets.

4. **Multilingual Support:** While most models are focused on English, a few (*e.g.*, MathGPT & Math-LLM) also support Chinese, showing a trend towards multilingual capabilities.

5. **Open Source:** A significant number of models are open-source, allowing broader access and fostering further research and development in the field.

In summary, the table reflects the rapid development of specialized Math-LLMs, with an increasing trend towards larger models, comprehensive evaluation benchmarks, and support for multilingual applications.

B Summary of Benchmarks

Table 3 summarizes the LLM-based benchmarks for mathematical reasoning.

C Illustration of More Cases

Figure 5 illustrates the diverse multimodal cases of mathematical reasoning settings.

C.1 Multimodal Plane Geometry Setting

The Multimodal Plane Geometry Setting involves mathematical problems that require understanding and reasoning about 2D geometric relationships. These problems typically focus on fundamental geometric concepts, such as points, lines, angles, and triangles, often leveraging trigonometric principles like sine, cosine, or tangent. Visually, these questions are characterized by clear plane diagrams with labeled points, angles, and lengths. Students need to interpret these visuals to solve for unknown distances, angles, or other parameters. The defining feature here is the emphasis on 2D spatial relationships and the need to derive solutions from diagrammatic representations that combine measurements and geometry.

C.2 Multimodal Solid Geometry Setting

The Multimodal Solid Geometry Setting shifts the focus from 2D to 3D shapes and figures, such as cylinders, spheres, cubes, or cones. These questions often require students to compute surface area, volume, or height based on given measurements or constraints. Visually, these questions feature 3D diagrams with dimensions like radius, height, or length, typically annotated on the figure to help guide problem-solving. The main distinction is

Benchmarks	Venue	Language	Size	Source	Level(s)	Evaluation	Model(s)	Task(s)
DynaMath (Zou et al., 2024) ★	ICLR'25	English	5,010	S P G	E M H U	Both	Closed/Open	S
MathCheck (Zhou et al., 2024d) ★	ICLR'25	English/Chinese	4,536	P	E M H U	Discriminative	Closed/Open/Math	S
GSM-Symbolic (Mirzadeh et al., 2024)	ICLR'25	English	5,000	P	E	Discriminative	Closed/Open	S
Omni-MATH (Gao et al., 2024)	ICLR'25	English	4,428	S	C	Discriminative	Closed/Open/Math	S
HARDMath (Fan et al., 2024a)	ICLR'25	English	1,466	G	U	Both	Closed/Open	S
OpenMathInstruct-2 (Toshniwal et al., 2024a)	ICLR'25	English	14,000,000	P G	E H C	Discriminative	Open/Math	S
UGMathBench (Xu et al., 2025b)	ICLR'25	English	5,062	S	U	Discriminative	Closed/Open/Math	S
M^3CoT_{math} (Chen et al., 2024b) ★	ACL'24	English	1,166	P G	C	Discriminative	Closed/Open	S
GSM-Plus (Li et al., 2024d)	ACL'24	English	10,552	P	E M H U	Generative	Closed/Open/Math	S
MuggleMath (Li et al., 2024c)	ACL'24	English	37,365	P	E H	Discriminative	Open	S
Olympiadbench (He et al., 2024) ★	ACL'24	English/Chinese	8,476	S	H C	Generative	Closed/Open/Math	S
MathBench (Liu et al., 2024b)	ACL Findings'24	English/Chinese	3,709	P S	E M H U	Generative	Closed/Open/Math	S
GeoEval (Zhang et al., 2024d) ★	ACL Findings'24	English	5,050	P G	E M H	Discriminative	Closed/Open/Math	S
QRData (Liu et al., 2024d)	ACL Findings'24	English	411	S	U	Discriminative	Closed/Open/Math	S
EIC-Math (Li et al., 2024e)	ACL Findings'24	English	1,800	P	E M H	Discriminative	Closed/Open	D O
Srivastava et al. (2024)	ACL Findings'24	English	-	P	H	Discriminative	Closed/Open	S
CHAMP (Mao et al., 2024)	ACL Findings'24	English	270	S	H	Generative	Closed/Open	S
IMO-AG-30 (Trinh et al., 2024)	Nature'24	English	30	S	C	Discriminative	Closed	P
PutnamBench (Tsoukalas et al.)	NeurIPS'24	English	1,697	S	C	Generative	Closed	S P
MATH-Vision (Wang et al., 2024a) ★	NeurIPS'24	English	3,040	S	E M H U	Discriminative	Closed/Open	S
CARP (Zhang et al., 2024a)	NeurIPS'24	Chinese	4,886	S	C	Discriminative	Closed	S
SMART-840 (Cherian et al., 2024) ★	NeurIPS'24	English	840	S	E M H	Discriminative	Closed/Open	S
OpenMathInstruct-1 (Toshniwal et al., 2024b)	NeurIPS'24	English	1,800,000	P	E M H C	Generative	Closed/Open/Math	S
Didolkar et al. (2024)	NeurIPS'24	English	8,600	P	E	Discriminative	Closed	S O
Putnam-AXIOM (Gulati et al., 2024)	NeurIPS Workshop'24	English	236	S	C	Discriminative	Closed/Open/Math	S
Scibench (Wang et al., 2023b) ★	ICML'24	English	869	S	U	Discriminative	Closed/Open	S
GeomVerse (Kazemi et al., 2023) ★	ICML Workshop'24	English	1,000	G	U	Discriminative	Closed	S
MathVista (Lu et al., 2023) ★	ICLR'24	English	6,141	S P	E M H U	Discriminative	Closed/Open	S
$MMMU_{math}$ (Yue et al., 2024a) ★	CVPR'24	English	540	S	U	Discriminative	Closed/Open	S
MathVerse (Zhang et al., 2024f) ★	ECCV'24	English	2,612	S P	H	Generative	Closed/Open	S
Mathador-LM (Kurtic et al., 2024)	EMNLP'24	English	-	G	E	Both	Closed/Open	S D
MM-MATH (Sun et al., 2024a) ★	EMNLP Findings'24	English	5,929	S	M H	Discriminative	Closed/Open	S D
Scieval (Sun et al., 2024b)	AAAI'24	English	15,901	S P	H	Both	Closed/Open	S
ArqMATH (Satpute et al., 2024)	SIGIR'24	English	450	P	U	Generative	Closed/Open/Math	S
IsoBench (Fu et al., 2024a) ★	COLM'24	English	1,887	S	E M H U	Discriminative	Closed/Open	S
$MMMU\text{-}Pro_{math}$ (Yue et al., 2024b) ★	arXiv'24	English	60	S	U	Discriminative	Closed/Open	S
MathOdyssey (Fang et al., 2024)	arXiv'24	English	387	S	H U C	Both	Closed/Open/Math	S
MathScape (Zhou et al., 2024b) ★	arXiv'24	Chinese	1,325	S	E M H	Generative	Closed/Open	S
U-Math (Chernyshev et al., 2024) ★	arXiv'24	English	1,100	S	U	Discriminative	Closed/Open/Math	S D
MathHay (Wang et al., 2024b)	arXiv'24	English	673	S P	H	Both	Closed/Open	S
ErrorRador (Yan et al., 2024a) ★	arXiv'24	English	2,500	S	E M H	Discriminative	Closed/Open	D
FaultyMath (Rahman et al., 2024) ★	arXiv'24	English	363	G	E M H	Discriminative	Closed/Open/Math	D
MathChat (Liang et al., 2024c)	arXiv'24	English	1,319	P	E	Both	Closed/Open/Math	S D O
E-GSM (Xu et al., 2024e)	arXiv'24	Chinese	4,500	P	E	Both	Closed/Open/Math	S O
Tangram (Tang et al., 2024a) ★	arXiv'24	English	4,320	S	E M H C	Discriminative	Closed/Open	O
CMM-Math (Liu et al., 2024c) ★	arXiv'24	Chinese	28,069	S	E M H	Both	Closed/Open/Math	S
CMMaTH (Li et al., 2024i) ★	arXiv'24	English/Chinese	23,856	S	E M H	Both	Closed/Open/Math	S
EAGLE (Li et al., 2024h) ★	arXiv'24	English	170,000	P	E M H	Discriminative	Closed/Open/Math	S
VisAidMath (Ma et al., 2024) ★	arXiv'24	English	1,200	S	M H C	Discriminative	Closed/Open	S
AutoGeo (Huang et al., 2024d) ★	arXiv'24	English	100,000	S	E M H U	Both	Closed/Open	O
NTKEval (Guo et al., 2024a)	arXiv'24	English	1,860	P G	H	Discriminative	Open	S
Mamo (Huang et al., 2024b)	arXiv'24	English	1,209	S G	U	Generative	Closed/Open/Math	O
RoMath (Cosma et al., 2024)	arXiv'24	Romanian	70,000	S	M H C	Discriminative	Closed/Open/Math	S
MaTT (Davoodi et al., 2024)	arXiv'24	English	1,958	S	U	Discriminative	Closed/Open	S
Li et al. (2024a)	arXiv'24	English	15,000	P	E M H	Generative	Closed/Open/Math	S
PolyMATH (Gupta et al., 2024) ★	arXiv'24	English	5,000	S	M H U	Discriminative	Closed/Open	S
SuperCLUE-Math6 (Xu et al., 2024b)	arXiv'24	English/Chinese	2,144	S	E	Generative	Closed/Open	S
TheoremQA (Chen et al., 2023)	EMNLP'23	English	800	S	U	Discriminative	Closed/Open	S
LILA (Mishra et al., 2022)	EMNLP'22	English	133,815	P	H	Discriminative	Closed	S
GeoQA (Chen et al., 2021) ★	ACL'21	Chinese	4,998	S	M	Discriminative	Open	S
MATH (Hendrycks et al., 2021)	NeurIPS'21	English	12,500	S	C	Discriminative	Closed	S

Table 3: **Overview of LLM-based benchmarks for mathematical reasoning**. ★refers to those designed to evaluate the multimodal mathematical setting. Different colors indicate different types for the following columns:
Source: S= Self-Sourced, P= Collected from Public Dataset, G= Generated by LLM
Level: E= Elementary, M= Middle School, H= High School, U= University, C= Competition, H= Hybrid
Task: S= Problem-Solving, D= Error Detection, P= Proving, O= Others

the incorporation of three-dimensional spatial reasoning and the need to analyze geometric properties of solids rather than flat, planar relationships. These tasks challenge students to bridge visual understanding with formulas involving multiple dimensions.

C.3 Multimodal Diagram Setting

In the Multimodal Diagram Setting, the problems revolve around interpreting visual data presented in the form of tables, charts, or diagrams. These tasks require students to extract numerical or categorical information and perform basic operations, such as addition, comparison, or selection. The visual components often include neatly organized tables, bar charts, or pie graphs, where the information is clearly labeled for accessibility. Unlike geometry-based problems, which require spatial reasoning, diagram settings focus on numerical literacy and the ability to synthesize information from structured visual data. This type highlights the integration of simple arithmetic and the comprehension of organized visual representations.

Math (M)LLMs	Organization	Release Date	Publication	Language	Parameter Size	Evaluation Benchmarks	Open Source
GPT-f (Polu and Sutskever, 2021)	OpenAI	Sep 2020	-	English	160M/400M/700M	-	✔
Hypertree Proof Search (Lample et al., 2022)	Meta	Nov 2022	NeurIPS'22	English	-	miniF2F/Metamath	-
Minerva (Lewkowycz et al., 2022)	Google	Jun 2022	NeurIPS'22	English	8B/62B/540B	MATH/MMLU-STEM/GSM8k	-
JiuZhang 1.0 (Zhao et al., 2022)	RUC & iFLYTEK	Jun 2022	KDD'22	English	145M	-	✔
GAIRMath-Abel (Chern et al., 2023)	Shanghai Jiaotong University	2023	-	English	7B/13B/70B	GSM8K/MATH/MMLU/SVAMP/SCQ5K-English/MathQA	✔
JiuZhang 2.0 (Zhao et al., 2023)	RUC & iFLYTEK	2023	KDD ADS'23	English	-	JCAG/JBAG (MathBERT/DART/JiuZhang)	✔
KwaiYiiMath (Fu et al., 2023)	Kuaishou	Jan 2023	-	English/Chinese	13B	GSM8K/CMath/KMath	-
MathCoder (Wang et al., 2023a)	CUHK	Jan 2023	ICLR'24	English	7B/13B	GSM8K/MATH	✔
Llemma (Azerbayev et al., 2023)	Princeton University & Eleuther AI	Jan 2023	-	English	7B/34B	MATH/GSM8k/MMLU-STEM/SAT/OCWCourse	✔
Skywork-13B-Math (Zeng et al., 2024) ★	SkyworkAI	Jan 2023	-	English	7B/13B	GSM8K/CMATH/MATH	✔
MathGPT (TALEducation, 2023)★	TAL Education Group	Aug 2023	-	English/Chinese	130B	CEval-Math/AGIEval-Math/APE5K/CMMLU-Math/GAOKAO-Math/Math401	-
WizardMath (Luo et al., 2023)	Microsoft	Aug 2023	ICLR'25	English	7B/70B	GSM8K/MATH	✔
MAmmoTH1 (Yue et al., 2023)	UWaterloo	Sep 2023	ICLR'24	English	7B/13B/70B	GSM/MATH/MMLU-STEM/AQuA/NumGLUE	✔
MathGLM (Yang et al., 2023b)	Tsinghua & Zhipu.AI	Sep 2023	-	English	10M/100M/500M/2B(Arith.)&335M/6B/10B (MWP)	BIG-bench/ Ape210K	✔
MetaMath (Yu et al., 2023)	Cambridge & Huawei	Sep 2023	-	English	7B/13B/70B	GSM8k/MATH	✔
DeepSeekMath (Shao et al., 2024)	DeepSeek AI	Jan 2024	-	English	7B	GSM8K/MATH/OCW/SAT/MMLU-STEM/CMATH/Gaokao-MathCloze/Gaokao-MathQA	✔
InternLM2.5-StepProver (Wu et al., 2024c)	Shanghai AI Lab	Jan 2024	-	English/Chinese	7B	miniF2F/Lean-Workbook-Plus/ProofNet/Putnam	✔
ChatGLM-Math (Xu et al., 2024f)	Zhipu.AI	Apr 2024	-	English/Chinese	32B	MathUserEval/Ape210k/CMath/GSM8k/MATH/Hungarian	-
Rho-Math (Lin et al., 2024)	Microsoft	Apr 2024	-	English	1B/7B	GSM8K/MATH/MMLU-STEM/SAT/SVAMP/ASDiv/MAWPS/TAB/MQA	✔
DeepSeekProver-V1 (Xin et al., 2024b)	DeepSeek AI	May 2024	-	English	7B	miniF2F/FIMO	-
InternLM2-Math (Wu et al., 2024c)	Shanghai AI Lab	May 2024	-	English/Chinese	1.8B/7B/20B/8x22B	MiniF2F-test/MATH/MATH-Python/GSM8K/MathBench-A/Hungary/	✔
JiuZhang 3.0 (Zhou et al., 2024a)	RUC & iFLYTEK	May 2024	NeurIPS'24	English	7B/8B	GSM8k/MATH/G-Hard/SVAMP/MAWPS/ASDiv/TabMWP	✔
MAmmoTH2 (Yue et al., 2024c)	UWaterloo	May 2024	-	English	7B/8B	TheoremQA/MATH/GSM8K/GPQA/MMLU-STEM/BBH	✔
Math-LLaVA (Shi et al., 2024)	NUS	Jun 2024	EMNLP Finding'24	English	13B	MMMU/MATH-V/MathVista	✔
Mathstral (MistralAI, 2024)	Mistral AI	Jul 2024	-	English	7B	MATH/GSM8K/GREMath/AMC2023/AIME2024/MathOdyssey	-
DeepSeek-Prover-V1.5 (Xin et al., 2024a)	DeepSeek AI	Aug 2024	-	English	7B	miniF2F-test/ProofNet	✔
Qwen2-Math (Qwen, 2024)	Alibaba	Aug 2024	-	English/Chinese	1.5B/7B/72B	GSM8K/Math/MMLU-STEM/CMATH/GaoKaoMath Cloze/-GaoKao Math QA	✔
Qwen2-Math-Instruct (Qwen, 2024)	Alibaba	Aug 2024	-	English/Chinese	1.5B/7B/72B	GSM8K/MATH/Minerva Math/GaoKao2023 En/Olympiad Bench/College Math/MMLU STEM/Gaokao/CMATH/CNMiddle School 24/AIME24/AMC23	✔
MathGLM-Vision (Yang et al., 2024b) ★	Tsinghua & Zhipu.AI	Sep 2024	-	English	9B/19B/32B	MathVista/MathVista(GPS)/MathVerse/Math-Vision/MMMU/MathVL	-
Math-LLM (Liu et al., 2024c) ★	East China Normal University	Sep 2024	-	Chinese	8.26B/7B/72B	CMM-Math/MathVista/Math-V	-
Qwen2.5-Math (Yang et al., 2024a)	Alibaba	Sep 2024	-	English/Chinese	1.5B/7B/72B	GSM8K/MATH/MMLU-STEM/CMATH/GaoKao Math	✔
Xwin-LM (Ni et al., 2024)	Microsoft	May 2024	-	English	7B/13B/70B	GSM8K/MATH	✔
MathCoder2 (Lu et al., 2024c)	CUHK	Nov 2024	ICLR'25	English	7B	GSM8K/MATH/SAT-Math/OCW/MMLU-Math	✔
math-specialized Gemini 1.5 Pro ★	Google	Not launched yet	-	English	-	MATH/AIME2024/Math Odyssey/HiddenMath/IMO Bench	-
k0-math (MoonshotAI, 2024)	Moonshot AI	Nov 2024	-	English/Chinese	-	KAOYAN/MATH/AIME/OMNI-MATH/GAOKAO/ZHONGKAO	-
Duolingo Math (Duolingo, 2024)	Duolingo	2024	-	English	-	-	-
Khanmigo (KhanAcademy, 2024)	Khan Academy	2024	-	English	-	-	-
Squirrel LAM (SquirrelAiLearning, 2024) ★	Squirrel Ai Learning	2024	-	Chinese	-	-	-

Table 4: **Overview of math-specific LLMs (sort by release date).** ★refers to those designed to support the multimodal mathematical setting.

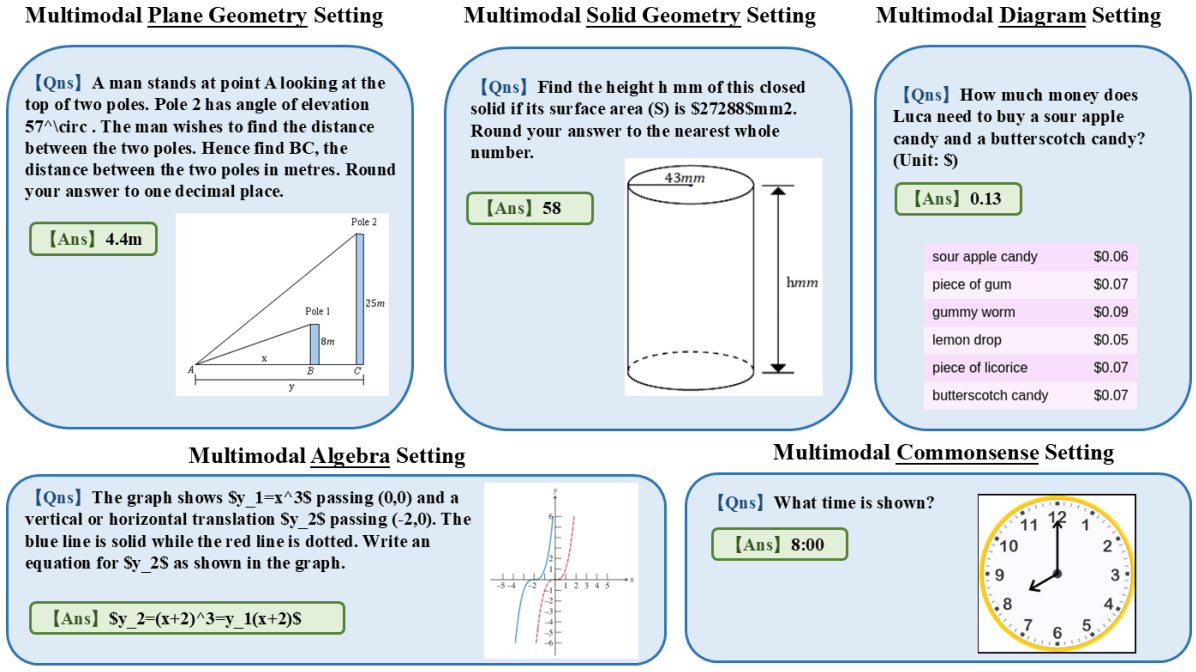

Figure 5: The illustration of diverse multimodal mathematical settings.

C.4 Multimodal Algebra Setting

The Multimodal Algebra Setting introduces problems that combine graphical representations and algebraic reasoning. These tasks often involve interpreting visual graphs, identifying equations, or understanding transformations such as translations or reflections. The visuals typically feature coordinate graphs with curves or lines, where solid and dotted lines may represent different functions or changes. Students are required to connect the visual graph to algebraic expressions, such as equations or transformations of functions. This type of question emphasizes the interplay between visual understanding (graph) and symbolic representation (algebra), making it distinct from purely numerical or geometric settings.

C.5 Multimodal Commonsense Setting

The Multimodal Commonsense Setting is characterized by problems that involve interpreting everyday visuals and applying logical reasoning. These ques-

tions present familiar objects, such as clocks, calendars, or real-world scenarios, where students must analyze the visual information to derive straightforward answers. Visually, these tasks feature clear and relatable imagery, like an analog clock with its hands pointing to a specific time. Unlike other types, commonsense settings rely less on abstract mathematical reasoning and more on practical interpretation of everyday visual cues. This setting highlights how mathematical understanding can intersect with routine, real-world observations.

C.6 Summary

In summary, the key differences among these types stem from their visual focus and cognitive demands. While plane and solid geometry emphasize spatial reasoning in 2D and 3D, respectively, diagram settings target numerical literacy through organized data. Algebra settings merge visual graphs with algebraic transformations, and commonsense settings leverage real-world visuals requiring practical logic. Each type uniquely integrates multimodal elements to challenge students across different mathematical skills.

D Details of Metrics

D.1 Discriminative Metrics

Discriminative tasks refer to evaluation processes where the outputs are typically binary, such as "Yes" or "No". These tasks often include multiple-choice questions, fill-in-the-blank problems, or judgment assessments. The evaluation metrics focus on LLM's accuracy in specific task types and its ability to control biases.

Accuracy (ACC): It measures the proportion of correctly predicted outcomes. The value should be as high as possible.

$$ACC = \frac{\sum_{1,m} x_i}{\sum_{1,n} y_j}$$

Where $\boldsymbol{x}_i$ represents the correct output for the $\boldsymbol{i}$-th instance, $\boldsymbol{y}_j$ represents the $\boldsymbol{j}$-th instance, m is the number of the correct instances and n is the number of the total instances.

Exact match: It evaluates the congruence between the answers generated by LLM and the correct ones. Specifically, in cases where the answer produced LLM coincides with the reference answer, a score of 1 point will be assigned. Conversely, if there is any discrepancy between them except for bias, a score of 0 point will be given.

F_1 score: It combines two crucial aspects, namely precision and recall, in order to comprehensively assess the accuracy of LLM. It is calculated as :

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The value of the F_1 score ranges from 0 to 1. A higher value of the F_1 score indicates better overall performance of LLM in terms of both precision and recall.

Macro-F_1 score: It calculates the F_1 score for each category separately and then takes the average of the F_1 scores of all categories, so as to obtain the overall performance of LLM on all categories.

Round-r accuracy: It is the proportion of correct answers given by a model on the question set Qr in round r. It is calculated as follows:

$$ACC_r(M) = \frac{\sum_{q \in Q_r} I[M(q) = g_t(q)]}{|Q_r|}$$

Here, $ACC_r(M)$ represents the accuracy of LLM M on question set Q_r in round r. I is an indicator function. When the answer $M(q)$ given by M for question q is consistent with the true answer $g_t(q)$ of the question, the value of I is 1; otherwise, it is 0. The symbol $\sum_{q \in Q_r}$ means summing over all questions in question set Q_r. $|Q_r|$ indicates the number of questions in question set Q_r.

ACC_{step} : It is used to evaluate LLM's ability to identify the first step where an error occurs. The accuracy for identifying the first erroneous step is calculated as follows:

$$ACC_{step} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S_{step,i} = G_{step,i})$$

Here, N is the total number of samples. For the i-th sample, $S_{step,i}$ is the predicted step where the error occurs, and $G_{step,i}$ is the ground truth label for the first erroneous step. The indicator function $\mathbb{I}(\cdot)$ returns 1 if the predicted step matches the ground truth and 0 otherwise.

ACC_{cate} : It is for assessing LLM's performance in categorizing the type of error. The accuracy for error categorization is defined by

$$ACC_{cate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(C_{error,i} = G_{error,i})$$

Here, N is the total number of samples. For the i-th sample, $C_{error,i}$ is the predicted error category, and $G_{error,i}$ is the ground truth label for the error category. The indicator function $\mathbb{I}(\cdot)$ has the same

meaning as in the previous metric, returning 1 if the predicted error category matches the ground truth and 0 otherwise.

The skill success rate: It measures the proportion of a model correctly applying major skills in problem-solving. It's calculated by analyzing test questions and determining correct use of major skills, then finding the ratio to total questions. For example, in triangle area calculation, checking use of the area formula. Similarly, **the secondary skill success rate** focuses on the proportion of correct application of secondary skills like understanding graphic properties and unit conversion, calculated by analyzing problem-solving and finding the ratio to total questions.

The False Positive Rate (FPR): It is the proportion of cases where the evaluation LLM misjudges an incorrect answer as a correct one. A low FPR indicates that LLM rarely misjudges incorrect student answers as correct.

The False Negative Rate (FNR): It is the proportion of cases where the evaluation LLM misjudges a correct answer as an incorrect one. A low FNR indicates that LLM is relatively accurate in correctly determining whether a student's answer is correct.

Mean Squared Error (MSE): It is a metric that measures the average of the squares of the differences between the LLM's predicted values and the actual true values. It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here, n represents the number of samples. For the i-th sample, y_i is the true value and $\hat{y}_i$ is the predicted value by LLM. The summation symbol $\sum_{i=1}^{n}$ means summing up the squared differences for all n samples. Dividing by n gives the average squared difference, which is the MSE. MSE should be as low as possible.

Average-Case Accuracy (A_{avg}): This metric evaluates the average accuracy of LLM across all variants of a seed question. It is calculated as the proportion of correct answers across all variants and seed questions. The formula is:

$$A_{avg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} I[\mathrm{Ans}(i,j) = \mathrm{GT}(i,j)]$$

where N is the total number of seed questions, M is the number of variants per seed question, and $I[\mathrm{Ans}(i,j) = \mathrm{GT}(i,j)]$ checks if the answer matches the ground truth.

Worst-Case Accuracy (A_{wst}): This evaluates the worst-case performance by considering the minimum accuracy across all variants of a seed question. It reflects the robustness of LLM against challenging variations. The formula is:

$$A_{wst} = \frac{1}{N}\sum_{i=1}^{N}\min_{j\in[1,M]} I[\mathrm{Ans}(i,j) = \mathrm{GT}(i,j)]$$

D.2 Generative Metrics

Generative tasks involve evaluating the content generated by LLM, typically encompassing free-form answers and responses to open-ended questions. These tasks focus primarily on assessing the extent of hallucinations in the generated content, especially when the content is not faithful to the given images. Evaluating generative tasks often requires more complex metrics, such as CHAIR and Faithscore, which measure hallucinations across different categories, including objects, attributes, and relationships within the generated content. These metrics provide a nuanced understanding of the fidelity and reliability of MLLMs in producing content aligned with the visual and textual inputs.

Reasoning Robustness (RR): This metric measures the relative robustness of LLM by comparing the worst-case performance to the average-case performance. The formula is:

$$RR = \frac{A_{wst}}{A_{avg}}$$

Repetition Consistency (RC): This evaluates the consistency of LLM's responses across repeated queries for the same question variant. It helps distinguish between variability due to randomness and systematic errors. The formula is:

$$RC(i,j) = \frac{1}{K}\sum_{k=1}^{K} I[\mathrm{Ans}_k(i,j) = \mathrm{Ans}(i,j)]$$

where K is the number of repetitions.

OpenCompass Scoring: It is a comprehensive evaluation framework that leverages the OpenCompass platform to assess the generative capabilities of LLM across multiple dimensions. Perplexity (PPL) evaluates the naturalness and fluency of generated text, with lower scores indicating greater model confidence and the ability to produce contextually coherent sequences. Simultaneously, CircularEval assesses the robustness and consistency of LLM in multiple-choice scenarios by evaluating its performance across N random permutations of

the options in an N-option question. A question is deemed correctly answered only if LLM provides the correct response for all permutations, highlighting its ability to handle randomized inputs reliably.

Bilingual Evaluation Understudy (BLEU): It evaluates the quality of text generation by measuring n-gram overlap between generated and reference texts, focusing on precision and brevity. Its formula is:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where BP is the brevity penalty, calculated as 1 if $c > r$, or $\exp(1 - r/c)$ if $c \leq r$, with c and r representing the lengths of the generated and reference texts, respectively. w_n denotes n-gram weights (typically uniform), and p_n is the precision of n-grams of size n. BLEU scores range from 0 to 1 (often expressed as percentages, 0-100%), with higher scores indicating greater similarity between the generated and reference texts.

Recall-Oriented Understudy for Gisting Evaluation-L (ROUGE-L): It evaluates the quality of generated text by measuring its similarity to reference text, focusing on sequence alignment and structural consistency through the Longest Common Subsequence (LCS). It calculates recall as the proportion of the LCS length relative to the reference text length. The formula of recall is:

$$R = \frac{\text{LCS}(\text{Generated, Reference})}{\text{Length}(\text{Reference})}$$

It also calculates precision as the proportion of the LCS length relative to the generated text length. The formula is:

$$P = \frac{\text{LCS}(\text{Generated, Reference})}{\text{Length}(\text{Generated})}$$

The F_1 score is a harmonic mean of precision and recall, expressed as:

$$F_1 = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

where β (commonly set to 1) controls the weighting of recall and precision. ROUGE-L scores range from 0 to 1, with higher scores indicating greater similarity between the generated and reference texts.

Consensus-based Image Description Evaluation (CIDEr): It is designed for image description tasks, measuring the semantic relevance of generated descriptions by calculating the TF-IDF weighted n-gram similarity with reference descriptions. The formula is:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_{j=1}^{m} \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \cdot \|g^n(s_{ij})\|}$$

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i)$$

Here, c_i is the candidate description, $S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$ is the set of reference descriptions, and m is the number of references. $g^n(c_i)$ and $g^n(s_{ij})$ are the TF-IDF weighted n-gram vectors for the candidate and reference descriptions, with $\|g^n(c_i)\|$ and $\|g^n(s_{ij})\|$ being their magnitudes. w_n is the weight for n-grams of different lengths, usually $w_n = 1/N$, where N is the maximum n-gram length. Scores range from 0 to 10, with higher scores indicating stronger alignment between candidate and reference descriptions.

Mathematical Symbol Similarity: This metric measures the similarity between the correct steps in a reasoning process and the steps generated by LLM, using symbolic computation software to perform the evaluation.

GPT Scoring: This metric evaluates the generated content based on scores assigned by GPT or other language models, focusing on the linguistic coherence and logical consistency of the text.

Context Length Generalization Efficacy (CoLeG-E): It is a metric used to measure LLM's consistency in answering variations of the same question across different context lengths. It is defined as:

$$CoLeG\text{-}E(M) = \frac{\sum_{q \in Q_R} \left[\bigwedge_{r=1}^{R} I[M(q^r) = gt(q^r)]\right]}{|Q_R|}$$

where Q_R represents the set of all questions under evaluation, and q^r refers to the r-th variation of a question q, corresponding to a specific context length. $M(q^r)$ is LLM's predicted answer for the r-th variation, while $M(q^r)$ denotes the ground truth answer. The indicator function $I[\cdot]$ equals 1 if LLM's answer matches the ground truth, and 0 otherwise. The logical AND operator $\Lambda_{r=1}^{R}$ ensures that the model must answer all variations of a question correctly for that question to be considered correctly answered.

Context Length Generalization Robustness (CoLeG-R): It measures LLM's robustness to context length expansion by quantifying the relative

drop in accuracy from initial to extended questions. It is defined as:

$$CoLeG\text{-}R(M) = 1 - \frac{\mathrm{ACC}_0(M) - \mathrm{ACC}_R(M)}{\mathrm{ACC}_0(M)}$$

Here, $\mathrm{ACC}_0(M)$ is the LLM's accuracy on the initial set of shorter-context questions Q_0, and $\mathrm{ACC}_R(M)$ is its accuracy on the extended longer-context questions Q_R. Higher CoLeG-R values indicate better robustness, with less performance degradation across context lengths.

Performance Drop Rate (PDR): This metric measures the relative decline in model performance when transitioning from the original dataset to the perturbed dataset. It is defined as:

$$\mathrm{PDR} = 1 - \frac{\sum_{(x,y)\in D_a} I[\mathrm{LLM}(x), y]/|D_a|}{\sum_{(x,y)\in D} I[\mathrm{LLM}(x), y]/|D|}$$

where D is the original dataset and D_a is the perturbed dataset. $I[\mathrm{LLM}(x), y]$ is an indicator function that checks if the LLM's output matches the ground truth y.

Accurately Solved Pairs (ASP): ASP measures the percentage of seed questions and their perturbed variations that are both correctly answered by LLM. It is defined as:

$$\mathrm{ASP} = \frac{\sum_{x,y;x',y'} I[\mathrm{LLM}(x), y] \cdot I[\mathrm{LLM}(x', y')]}{N \cdot |D|}$$

where x and x' are a seed question and its variation, respectively. N is the number of perturbations per question. $|D|$ is the total number of seed questions.

Mean Average Precision (mAP): It is a metric that evaluates LLM's ability to rank relevant answers higher in its output list for a given query. It is defined as:

$$mAP = \frac{1}{|Q|} \sum_{q\in Q} AP(q)$$

$$AP(q) = \frac{1}{m} \sum_{k=1}^{m} P(k)$$

$$P(k) = \frac{\text{\# relevant ans retrieved up to position } k}{k}$$

Here, Q represents the set of all queries in the dataset. $AP(q)$ is the Average Precision for query q, calculated as the mean of the precision values $P(k)$ at ranks where relevant answers appear.$P(k)$ is the precision at rank k, representing the proportion of relevant answers retrieved up to position k. m is the total number of relevant answers for query q.

Training Set Coverage (TSC): It measures how effectively LLM has learned to generate correct solutions for tasks similar to those in its training set. TSC is particularly useful in cross-domain or cross-modal tasks, where it assesses LLM's ability to generalize learned patterns to problems aligned with its training data. Higher TSC scores indicate better learning and consistency, while lower scores suggest insufficient training or overfitting.

Pass@N: This metric measures the likelihood of LLM generating at least one correct solution within N attempts for a given problem. Formally:

$$Pass@N = \mathbb{E}_{\mathrm{Problems}}[\min(c, 1)]$$

where c represents the number of correct answers out of N responses. A higher Pass@N indicates a greater chance of producing a correct answer in multiple attempts, reflecting LLM's potential capability.

PassRatio@N: This metric calculates the proportion of correct answers among N generated responses for a given problem. It is defined as:

$$PassRatio@N = \mathbb{E}_{\mathrm{Problems}}\left[\frac{c}{N}\right]$$

where c is the count of correct answers. This metric reflects LLM's stability in consistently generating correct answers. It can be considered analogous to Pass@1 but offers reduced variance.

E Summary of Methods

Table 5 summarizes the LLM-based methods for mathematical reasoning.

F More Details of Challenges

F.1 Discussion of Data Bottlenecks

We dive into the three bottlenecks of multimodal mathematical datasets as follows.

❶ Bottleneck in Data Quality:

1. **Labeling Noise and Modality Alignment**: Multimodal math problems often involve complex associations between text, formulas, and charts. Mismatches between text descriptions and images/formulas (*e.g.,* incorrect axis labels, contradictions between geometry figures and problem statements) can severely impair the model's ability to understand cross-modal relationships.

Methods	Venue	Evaluated Math Dataset(s)	Task(s)	Scope(s)	LLM as Enhancer	LLM as Reasoner	LLM as Planner
MAVIS (Zhang et al., 2024g) ★	ICLR'25	MathVerse/GeoQA/MathVista/MMMU/MathVision	S	M	✔	✔	
TVM (Lee et al., 2024)	ICLR'25	GSM8K/MATH	S	A		✔	
MathCoder2 (Lu et al., 2024c)	ICLR'25	GSM8K/MATH/SAT-Math/OCW/MMLU-Math	S P	M	✔	✔	
Xiong et al. (2024)	ICLR'25	GSM8K/MATH	S P	A		✔	✔
TSMC (Feng et al., 2024)	ICLR'25	GSM8K/MATH500	S P	A		✔	
AlphaGeometry (Trinh et al., 2024) ★	Nature'24	IMO-AG-30	S P	G	✔	✔	
Masked Thought (Chen et al., 2024a)	ACL'24	GSM8K/MATH/GSM8K-RFT/MetaMathQA/MathInstruct	S	A	✔	✔	
MathGenie (Lu et al., 2024b)	ACL'24	GSM8K/MATH/SVAMP/Simuleq/Mathematics	S	A	✔	✔	
MATH-SHEPHERD (Wang et al., 2024c)	ACL'24	GSM8K/MATH	S P	A	✔	✔	
SEGO (Zhao et al., 2024)	ACL'24	GSM8K/MATH	S P	A	✔	✔	
Deng et al. (2023)	ACL Workshop'24	GSM8K/SVAMP/MultiArith/MathQA/CSQA	S	A		✔	
MathCoder (Wang et al., 2023a)	ICLR'24	GSM8K/MATH	S P	A		✔	
ToRA (Gou et al., 2023)	ICLR'24	GSM8K/MATH	S P	A		✔	✔
Visual Sketchpad (Hu et al., 2024) ★	NeurIPS'24	Geometry3K/ IsoBench	S	G			✔
JiuZhang 3.0 (Zhou et al., 2024a)	NeurIPS'24	GSM8K/MATH/SVAMP/ASDiv/MAWPS/CARP	S P	A	✔	✔	
Minimo (Poesia et al., 2024)	NeurIPS'24	-	P	A		✔	
DART-Math (Tong et al., 2024)	NeurIPS'24	MATH/GSM8K/College/DM/Olympiad/Theorem	S P	A	✔	✔	
Li et al. (2024f)	NeurIPS'24	GSM8K/MATH	S	A	✔		
MACM (Lei et al., 2024)	NeurIPS'24	MATH	S	A		✔	
Sinha et al. (2024) ★	NeurIPS Workshop'24	IMO-AG-30	S P	G		✔	
SBIRAG (Dixit and Oates, 2024)	NeurIPS Workshop'24	GSM8K	S	A		✔	
MathScale (Tang et al., 2024b)	ICML'24	-	-	-		-	
VerityMath (Han et al., 2023)	ICML Workshop'24	GSM8K	S	A		✔	
RefAug (Zhang et al., 2024j)	EMNLP'24	GSM8K/MATH/Mathematics/MAWPS/SVAMP/MMLU-Math/SAT-Math/MathChat-FQA/MathChat-EC/MINI-Math	S D P	A	✔	✔	
Math-LLaVA (Shi et al., 2024) ★	EMNLP Findings'24	MathVista/Math-V	S P	M	✔	✔	
COPRA (Thakur et al., 2024)	COLM'24	miniF2F-test	S	A			✔
PRP (Wu et al., 2024b)	AAAI'24	MAWPS/ASDivA/Math23k/SVAMP/UnbiasedMWP	S	A		✔	
PERC (Jin et al., 2024)	L@S'24	PERC	S	A		✔	
Math-PUMA (Zhuang et al., 2024) ★	arXiv'24	MathVerse/MathVista/WE-MATH	S P	M		✔	
MultiMath (Peng et al., 2024) ★	arXiv'24	MathVista/MathVerse/MultiMath-300K	S P	M		✔	
MathAttack (Zhou et al., 2024e)	arXiv'24	GSM8K/MultiAirth	S	A		✔	
MinT (Liang et al., 2023b)	arXiv'24	GSM8K/MathQA/CM17k/Ape210k	S	A		✔	
DotaMath (Li et al., 2024b)	arXiv'24	GSM8K/MATH/Mathematics/SVAMP/TabMWP/ASDiv	S	A	✔	✔	
DFE-GPS (Zhang et al., 2024i)	arXiv'24	FORMALGEO7k	S	G	✔	✔	
PGPSNet-v2 (Zhang et al., 2024e) ★	arXiv'24	Geometry3K/PGPS9K	S	G D	✔	✔	
LLaMA-Berry (Zhang et al., 2024b)	arXiv'24	GSM8K/MATH/GaoKao2023En/OlympiadBench/College Math/MMLU STEM	S P	A		✔	
Skywork-Math (Zeng et al., 2024) ★	arXiv'24	GSM8K/MATH	S P	A	✔	✔	
SIaM (Yu et al., 2024a)	arXiv'24	GSM8K/CMATH	S P	A		✔	
InternLM-Math (Ying et al., 2024)	arXiv'24	GSM8K/MATH	S P	A		✔	
MathGLM-Vision (Yang et al., 2024b) ★	arXiv'24	MathVista/MathVerse/MathVision	S P	M	✔	✔	
Qwen2.5-Math (Yang et al., 2024a) ★	arXiv'24	GSM8K/MATH/MMLU-STEM/CMATH/GaoKao-Math-Cloze/GaoKao-Math-QA	S P	A	✔	✔	
S3c-Math (Yan et al., 2024c)	arXiv'24	GSM8K/MATH/SVAMP/Mathematics	S P	A	✔	✔	
SIRP (Wu et al., 2024a)	arXiv'24	CSQA/GSM8K/MATH/MBPP	S P	A		✔	
AIPS (Wei et al., 2024)	arXiv'24	MO-INT-20	S	G		✔	
DeepSeekMath (Shao et al., 2024)	arXiv'24	GSM8K/MATH/OCW/SAT/MMLU STEM/CMATH/-Gaokao MathCloze/Gaokao MathQA	S	A		✔	
MMIQC (Liu et al., 2024a)	arXiv'24	MATH/MMIQC	S	A	✔	✔	
LANS (Li et al., 2023c) ★	arXiv'24	Geometry3K/PGPS9K	S	G D		✔	
VCAR (Jia et al., 2024) ★	arXiv'24	MathVista/MathVerse	S	M		✔	
KPDDS (Huang et al., 2024c)	arXiv'24	GSM8k/MATH/SVAMP/TabMWP/ASDiv/MAWPS	S	A	✔	✔	
HGR (Huang et al., 2024a) ★	arXiv'24	Geometry3K	S	G	✔	✔	
InfiMM-Math (Han et al., 2024) ★	arXiv'24	GSM8K/MMLU/MathVerse/We-Math	S	A	✔		
CoSC (Han et al., 2024)	arXiv'24	GSM8K/MATH	S P	A	✔	✔	
SICCV (Liang et al., 2024b)	arXiv'24	GSM8k/MATH500	S P	A		✔	
BEATS (Sun et al., 2024c)	arXiv'24	GSM8K/MATH/SVAMP/SimulEq/NumGLUE	S P	A		✔	
MindStar (Kang et al., 2024)	arXiv'24	GSM8K/MATH	S P	A		✔	
UMM (Zhang et al., 2024h)	arXiv'24	MMLU/GSM8K-COT/GSM8K-Coding/MATH-COT/MATH-Coding/HumanEval/InfiBench	S P	A		✔	
STIC (Deng et al., 2024) ★	arXiv'24	ScienceQA/TextVQA/ChartQA/LLaVA-Bench/MMBench/MM-Vet/MathVista	S	M		✔	
SPMWPs (Zhang et al., 2023)	ACL'23	GSM8K	S	A		✔	
CoRe (Zhu et al., 2022)	ACL'23	GSM8K/ASDiv-A/SingleOp/SinlgeEq/MultiArith	S	A		✔	
TabMWP (Lu et al., 2022a)	ICLR'23	TabMWP	S	A D		✔	
Chameleon (Lu et al., 2024a) ★	NeurIPS'23	ScienceQA/TabMWP	S	A D			✔
ATHENA (Kim et al., 2023)	EMNLP'23	MAWPS/ASDivA/Math23k/SVAMP/UnbiasedMWP	S P	A		✔	
UniMath (Liang et al., 2023a) ★	EMNLP'23	SVAMP/GeoQA/TabMWP/MathQA/UniGeo-Proving	S P	A D		✔	
Jiuzhang 2.0 (Zhao et al., 2023)	KDD'23	MCQ/BFQ/CAG/BAG/KPC/QRC/JCAG/JBAG	S	A		✔	
TCDP (Qin et al., 2023)	TNNLS'23	Math23k/CM17K	S	A		✔	
UniGeo (Chen et al., 2022) ★	EMNLP'22	GeoQA/UniGeo	S P	G		✔	
LogicSolver (Yang et al., 2022)	EMNLP Findings'22	InterMWP/Math23K	S	A	✔	✔	
Jiuzhang (Zhao et al., 2022)	KDD'22	KPC/QRC/QAM/SQR/QAR/MCQ/BFQ/CAG/BAG	S	A		✔	
MWP-BERT (Liang et al., 2021)	NAACL'22	Math23k/MathQA/Ape-210k	S	A		✔	
Inter-GPS (Lu et al., 2021) ★	ACL'21	Geometry3K/GEOS	S	G		✔	

Table 5: **Overview of LLM-based methods for mathematical reasoning**. ★refers to those specifically designed to tackle the multimodal mathematical setting. Different colors indicate different types for the following columns:
Task: S= Problem-Solving, D= Error Detection, P= Proving, O= Others
Scope: G= Geometry, A= Algebra, D= Diagram, M= General Math

2. **Lack of Deep Annotation for Problem Solving Process**: Most datasets only provide final answers, lacking intermediate steps such as algebraic transformations or construction of geometric auxiliary lines, making it difficult for models to learn the mathematical thinking chain (Chain-of-Thought).

❷ Bottleneck in Data Diversity:

1. Limited Coverage of Problem Types and Scenarios: Existing datasets are mostly focused on basic math areas (*e.g.,* algebraic equations, simple geometry) and insufficiently cover higher-level math (*e.g.,* topology, discrete mathematics) or real-world scenarios (*e.g.,* physics modeling, financial calculations).

2. Monotony in Multimodal Combination Patterns: Modal interactions are often simple con-

catenations (*e.g.,* text + static charts) without dynamic interactions (*e.g.,* scalable geometric figures), or multi-step cross-modal reasoning (*e.g.,* generating charts from text descriptions and then solving problems).

❸ **Bottleneck in Data Scale:**

1. **High Cost of High-Quality Data Acquisition**: Mathematical problems need to be designed by experts and ensure multimodal consistency, which leads to long production cycles and high costs. Additionally, there is data scarcity for long-tail problems (*e.g.,* niche branches of mathematics), which cannot be supplemented by scraping existing resources (*e.g.,* textbooks, online question banks).

2. **Imbalance in Modal Data Volumes**: Text data volumes far exceed those of image/symbol modalities, leading to models' insufficient feature extraction capability for non-text modalities.

❹ Based on recent trends in the latest works, we further propose the following **actionable suggestions to address these dataset bottlenecks**:

1. **Innovation in Data Generation Techniques**: Combine formal mathematical engines (*e.g.,* Lean, Coq) to generate verifiable reasoning steps, use programmatic rendering tools (*e.g.,* TikZ, GeoGebra) to automatically generate precise charts, and design semi-automated annotation pipelines that reduce manual labor through large models generating drafts and experts refining them.

2. **Diversity Enhancement Strategies**: Construct interdisciplinary, cross-cultural benchmark datasets (*e.g.,* math-physics cross-domain problems), utilize crowdsourcing platforms to collect real-world scenario problems, and explore controllable data augmentation techniques, such as rule-based problem deformation (*e.g.,* modifying parameters or replacing chart elements).

3. **Scaling and Resource Integration**: Encourage collaborative dataset creation within the academic community (similar to ProofWiki), integrate existing educational resources (*e.g.,* Khan Academy video-text analysis), and use synthetic data to fill long-tail gaps while improving model robustness to synthetic noise through adversarial training.

F.2 Limited Domain Generalization in Multimodal Contexts

We further discuss the challenge of *limited domain generalization* in multimodal contexts through the perspective of the methodology paradigm.

1. **LLM as Enhancer**: Generate mixed-domain problems (*e.g.,* combining algebraic equations with geometric figures) to force the model to learn cross-domain associations. Explicitly add domain labels (*e.g.,* "spatial reasoning" label for geometry problems) to guide the model in distinguishing domain-specific features. The limitation of this paradigm is that enhanced data may lack the real-world complexity of domain intersections.

2. **LLM as Reasoner**: Fine-tune the model separately for different mathematical domains (*e.g.,* algebra, geometry) to learn domain-specific visual patterns (*e.g.,* encoding geometric properties in figures). Use domain-specific few-shot examples (*e.g.,* providing figure-text associations in geometry) to guide the model in switching reasoning modes. The limitation is that the model's capacity may be limited, making it difficult to master multiple significantly different domains simultaneously (*e.g.,* switching from algebraic symbol manipulation to geometric spatial reasoning).

3. **LLM as Planner**: Based on the problem domain (*e.g.,* detecting the "triangle" keyword), call specialized tools (*e.g.,* geometric theorem prover). For composite problems (*e.g.,* algebraic-geometry equations), coordinate symbolic computation tools (*e.g.,* Mathematica) and graphical reasoning tools (*e.g.,* GeoGebra). The limitation is that domain boundary issues (*e.g.,* math word problems requiring commonsense reasoning) may fail to route to the appropriate tools.

F.3 Error Feedback Limitations in Multimodal Contexts

We further discuss the challenge of *error feedback limitations* in multimodal contexts through the perspective of the methodology paradigm.

1. **LLM as Enhancer**: Inject cross-modal errors (e.g., plot errors in function curves while the text description is correct) to train the model to detect contradictions. The limitation is that

labeling error types is costly and it's difficult to cover all long-tail errors.

2. **LLM as Reasoner**: Decompose reasoning into "computation-logic-conclusion" stages and cross-check text derivations with graphical information (*e.g.,* verify function extrema calculations using coordinates in the image). The limitation is that self-doubt relies on the model's prior knowledge of error types, potentially missing rare error patterns in the training data.

3. **LLM as Planner**: Use OCR tools to extract symbols from figures and compare them with the text description to detect misunderstandings. The limitation is that tool invocation delays affect real-time performance, and some errors require manually defined detection rules.

F.4 How Test-Time Scaling Techniques Handle Other Challenges

We believe that test-time scaling techniques (Xu et al., 2025a; Li et al., 2025; Besta et al., 2025; Muennighoff et al., 2025; Chen et al., 2024c, 2025) can also help handle other challenges discussed in Section 4, especially the following three challenges.

❶ Insufficient Visual Reasoning:

1. **Enhancement of Multimodal Reasoning Chains**: During reasoning, generate multi-step visual-symbol joint inference paths. For example, using CoT prompts to guide the model in decomposing geometric shapes into angle, side length, and other symbolic constraints, and then calling a geometry solver to validate spatial relationships.

2. **Visual-Symbol Alignment Verification**: Use Best-of-N sampling to generate multiple candidate diagram parsing results and call external OCR tools or geometry validators (*e.g.,* GeoGebra) to detect visual-text consistency and filter out erroneous explanations.

3. **Limitations**: Parsing complex visual details (*e.g.,* topological structures) depends on the pretrained visual encoder's capabilities. If the training data coverage is insufficient, test-time strategies may not be able to compensate.

❷ Limited Domain Generalization:

1. Dynamic Domain Routing: Use Beam Search Process Reward Model (PRM) to select domain-specific inference paths based on problem types (*e.g.,* detecting the "triangle" keyword and choosing between algebra solvers or geometry theorem provers).

2. Meta-learning Optimization: Fine-tune the model on a small number of domain-specific samples via Test-Time Training (TTT) to quickly adapt to new domains (*e.g.,* probability and statistics problems)

3. Limitations: Problems with blurred domain boundaries (*e.g.,* math application problems involving common sense reasoning) may fail due to routing errors.

❸ Error Feedback Limitations:

1. Process Supervision Reinforcement: Use PRM to validate each step of reasoning in real-time. If an error is detected (*e.g.,* misuse of integration symbols), backtrack and correct the path; combine Self-Consistency by generating multiple inference paths and selecting the one with no contradictions via majority voting.

2. Limitations: The reliability of PRM depends on the coverage of error types in the training data. Long-tail errors such as rare symbol confusions may be overlooked.

In summary, combining the flexibility of test-time scaling with the specialization of multimodal tools can help mitigate the core challenges in multimodal mathematical reasoning. However, **it is crucial to balance computational efficiency and accuracy**.