Position: Multimodal Large Language Models Can Significantly Advance Scientific Reasoning

Anonymous ACL submission

Abstract

Scientific reasoning, the process through which humans apply logic, evidence, and critical thinking to explore and interpret scientific phenomena, is essential in advancing knowledge reasoning across diverse fields. However, despite significant progress, current scientific reasoning models still *struggle with generalization across domains and often fall short of multimodal perception*. Multimodal Large Language Models (MLLMs), which integrate text, images, and other modalities, present an exciting opportunity to overcome these limitations and enhance scientific reasoning. Therefore, **this position paper argues that MLLMs can significantly advance scientific reasoning** across disciplines such as mathematics, physics, chemistry, and biology. First, we propose a four-stage research roadmap of scientific reasoning capabilities, and highlight the current state of MLLM applications in scientific reasoning, noting their ability to integrate and reason over diverse data types. Second, we summarize the key challenges that remain obstacles to achieving MLLM's full potential. To address these challenges, we propose actionable insights and suggestions for the future. Overall, our work offers a novel perspective on MLLM integration with scientific reasoning, providing the LLM community with a valuable vision for achieving Artificial General Intelligence (AGI).

1 Introduction

Scientific reasoning, at its core, is the process through which humans apply logic, evidence, and critical thinking to explore and interpret phenomena in various scientific domains (Bao et al., 2009; Lawson, 2004). This cognitive ability is essential not only for advancing knowledge but also for fostering a deeper understanding of the natural world, particularly in fields such as mathematics, physics, chemistry, and biology. In education, medicine, finance, AI for Science, and other domains, scientific reasoning serves as a cornerstone for cultivating problem-solving skills, analytical thinking, and innovation (Huang and Chang, 2022; Willemsen et al., 2023; Qian et al., 2025; Jiang et al., 2024a; Jadon et al., 2025). However, despite shared objectives, each domain has unique characteristics in terms of data representation, knowledge construction, and reasoning methods (Ferrag et al., 2025; Chen et al., 2025).

In response to these challenges, the scientific community has explored a range of approaches, from traditional statistical methods to the more recent advancements in deep learning, with the goal of improving knowledge reasoning across disciplines (Goodman, 2016; Lu et al., 2022b). While significant progress has been made in enhancing scientific reasoning within specific domains, a gap remains in the broader context of scientific research. Current scientific reasoning models, and even those targeted toward domain-specific applications, are still *far from achieving the generalization capabilities necessary for Artificial General Intelligence (AGI), which aims to exhibit unified reasoning across all fields* (Birhane et al., 2023).

The rapid rise of Large Language Models (LLMs) in recent years has brought transformative changes across various domains, pushing the boundaries of what is possible in natural language processing and understanding (Min et al., 2023; Zhao et al., 2023a). Despite their remarkable zero-shot reasoning abilities, many areas, particularly in scientific fields, require multimodal inputs to build a comprehensive understanding of knowledge. This has led to the emergence and growth of Multimodal Large Language Models (MLLMs), which are capable of integrating and reasoning over multiple types of data, such as text, images, and other modalities (Liang et al., 2024b; Bai et al., 2024a). MLLMs are not only revolutionizing language understanding but also paving the

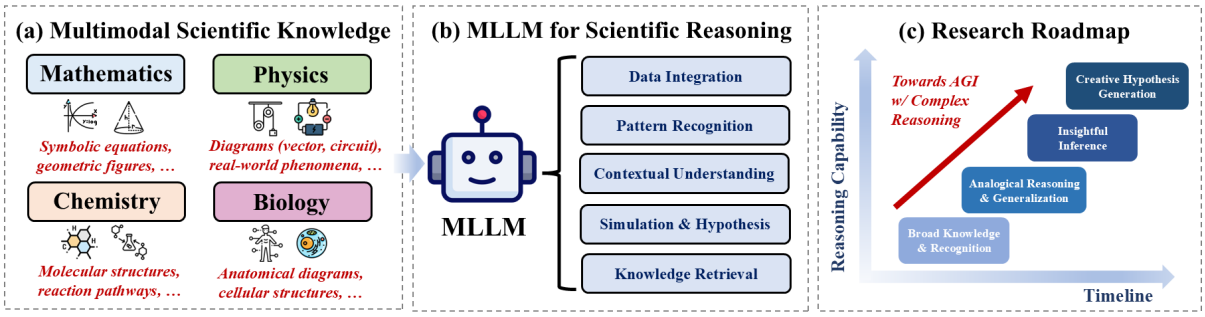

Figure 1: The big picture of our position. We focus on multimodal scientific fields, especially mathematics, physics, chemistry, and biology as our scope (a), and we advocate leveraging MLLMs with multiple reasoning functions for scientific reasoning (b). We further propose a four-stage roadmap for scientific reasoning capability, ultimately achieving AGI (c).

way for advancements in scientific reasoning by processing complex multimodal data in ways that were previously unachievable (Zou et al., 2025; Ke et al., 2025; Xu et al., 2025; Zha et al., 2025).

Building on the rapid evolution of MLLMs and the growing demand for enhanced multimodal reasoning capabilities in scientific domains (Reddy and Shojaee, 2024; Zhang et al., 2024d), **this paper proposes a position: MLLMs can significantly advance scientific reasoning**. By integrating multimodal learning techniques, MLLMs have the potential to address the pressing challenges in scientific reasoning, as shown in Figure 1. This paper aims to break down the analysis into three key sections: the current state of MLLMs in scientific reasoning, the challenges encountered, and the future steps needed to achieve greater success.

Our analysis delves into these topics in detail, starting with an overview of background and research roadmap (Section 2), followed by how MLLMs are currently handling scientific reasoning (Section 3) as well as a discussion of the challenges faced (Section 4), and concluding with the promising opportunities that lie ahead (Section 5). In particular, we first explore how MLLMs have been applied across scientific disciplines, detailing the techniques used for data integration, pattern recognition, contextual understanding, and more. The Challenges section examines the inherent difficulties, ranging from technical to ethical aspects, that need to be overcome for MLLMs to achieve optimal performance. Finally, in the Discussion section, we highlight future steps with relevant literature, from data & training strategies to agent-based collaboration, as key areas for advancing the integration of MLLMs with scientific reasoning. Through this, we aim to provide a comprehensive understanding of the landscape and offer strategic insights for future research in the intersection of MLLMs and scientific reasoning[1].

Our position and analytical framework contribute in three significant ways. ❶ We present **a novel perspective** on the integration of MLLMs with scientific reasoning, highlighting how this synthesis could reshape research in the field. ❷ We offer **a systematic review and categorization** of recent advancements, showcasing representative work and outlining a clear roadmap for future progress. ❸ We identify and explore **future opportunities**, providing the community with promising insights that could guide the next generation of research in scientific reasoning.

2 Background

2.1 Scientific Reasoning

Scientific reasoning is the intellectual process of forming hypotheses, interpreting evidence, and applying logical frameworks to solve problems or explain phenomena (Bao et al., 2009; Lawson, 2004). Its importance spans diverse scientific domains, such as mathematics, physics, chemistry, and biology, where it drives discovery, fosters understanding, and enables practical innovation. With the rise of multimodal data, scientific reasoning increasingly requires integrating and synthesizing information from multiple sources, including textual, visual, and other modalities[2].

The significance of scientific reasoning in the MLLM context is profound. By enabling models to connect disparate data points and infer relationships across modalities, MLLMs hold the potential to transform how researchers approach interdisciplinary problems. This capability is critical for addressing grand challenges such as climate modeling, drug discovery, *etc* (Zhang et al.,

[1] See more clarification of our scope in Appendix A.
[2] See the formal formulation of the task in Appendix B.

2024d). Moreover, enhancing MLLM-based scientific reasoning aligns with broader goal of advancing AGI, as it exemplifies synthesis of learning, abstraction, and decision-making across domains (Guo et al., 2025a; Thawakar et al., 2025).

2.2 Multimodal Large Language Models

Most existing MLLMs consist of three primary modules: a modality encoder, an LLM module, and a projector between them (Fu et al., 2024). Typically, the modality encoder extracts embeddings from non-language modalities such as images or audio, which are then projected into the word space of the LLM via the projector. The post-projection embeddings are subsequently combined with word embeddings derived from system prompts and user queries to serve as input for the LLM. Similar to LLMs, MLLMs generate responses in an autoregressive manner:

$$p(\mathbf{w}_O \mid \mathbf{w}_V, \mathbf{w}_T) \sim \prod_{t=1}^{L} P(w_t \mid w_{<t}, \mathbf{w}_V, \mathbf{w}_T)$$

Here, $\mathbf{w}_V$ and $\mathbf{w}_T$ denote the post-projection embeddings and word embeddings respectively, while $\mathbf{w}_O = \{w_{o,t}\}_{t=1}^{L}$ represents the generated word token sequence of length L. With their capability to comprehend visual inputs, contemporary MLLMs demonstrate remarkable performance in various tasks, including visual question answering (VQA) (Ishmam et al., 2024; Uppal et al., 2022; Dang et al., 2024a), image captioning (Vaishnavi and Narmatha, 2024; Agarwal and Verma, 2024), and multimodal reasoning (Yan and Lee, 2024; Yan et al., 2024a; Huo et al., 2024).

Presently, there is a plethora of open-source foundation MLLMs capable of general multimodal tasks. Notable examples include LLaVA family (Liu et al., 2023b,a, 2024b), Qwen-VL series (Bai et al., 2023; Wang et al., 2024f), InternVL series (Chen et al., 2024b,a), LLaMA-3.2-Vision (Dubey et al., 2024), etc. Despite these advancements, open-source MLLMs still lag behind closed-source models like GPT-4o (Achiam et al., 2024), Claude (Anthropic, 2024), and Gemini-Pro (Team et al., 2024a) in complex reasoning tasks (Liu et al., 2025d; Yue et al., 2024a). With the emergence of o1-like reasoning models (Jaech et al., 2024; Zeng et al., 2025; Liu et al., 2025b), preliminary efforts are underway to elicit the slow-thinking capabilities of MLLMs, as seen in works like QvQ (Qwen, 2024a), Mulberry (Yao et al., 2024), Virgo (Du et al., 2025), *etc*.

3 How MLLMs Benefit Scientific Reasoning

3.1 Research Roadmap

The development of (M)LLMs for scientific reasoning can be categorized into four progressive stages: *Broad Knowledge and Recognition*, *Analogical Reasoning and Generalization*, *Insightful Inference*, and *Creative Hypothesis Generation*. Each stage is defined by its unique characteristics across four dimensions: data and knowledge requirements, reasoning mechanisms, model generalization, and applications and impact (See the detailed comparisons among the four dimensions in Appendix C). Figure 1 (c) provides an overview of four stages, highlighting their evolution progress.

Stage 1: Broad Knowledge and Recognition. The initial stage focuses on building a strong foundational understanding across domains. MLLMs in this stage rely on highly diverse and multimodal datasets to capture a broad range of knowledge. Reasoning mechanisms are primarily retrieval-based, with emphasis on pattern recognition, data alignment, and summarization. Model generalization remains limited, operating primarily within predefined domains (White, 2023; Pei et al., 2024; Chen et al., 2023c).

Stage 2: Analogical Reasoning and Generalization. This stage emphasizes the ability to draw connections and analogies across domains. Data requirements shift towards moderately diverse datasets that emphasize relationships and cross-domain patterns. Reasoning mechanisms incorporate relational reasoning and analogical thinking, enabling MLLMs to generalize effectively across domains. Applications include interdisciplinary problem-solving, transfer learning, and identifying cross-domain insight, reflecting a moderate increase in complexity and impact (Webb et al., 2023; Lewis and Mitchell, 2024).

Stage 3: Insightful Inference. The third stage focuses on inferring deep insights from minimal and high-context data. Data requirements narrow to low-diversity, domain-specific datasets, allowing MLLMs to develop nuanced understanding. Reasoning mechanisms involve predictive reasoning and contextual interpretation, enabling the model to deduce complex outcomes. Generalization becomes highly context-specific, and applications include optimization and predictive modeling, making this stage highly impactful (Melko and Carrasquilla, 2024; Barman et al., 2025).

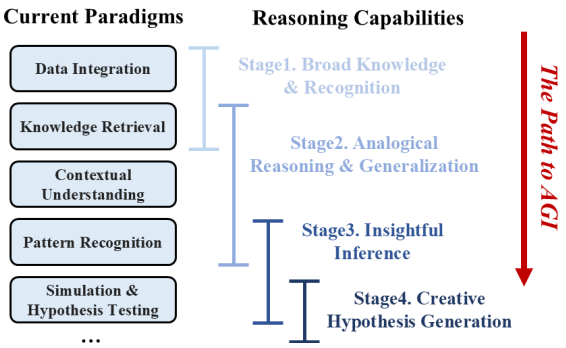

Figure 2: Overview of MLLM-based scientific reasoning paradigms and corresponding reasoning capabilities.

Stage 4: Creative Hypothesis Generation. In the final stage, MLLMs achieve the ability to generate innovative hypotheses and explore uncharted territories. Data requirements include highly diverse and synthetic datasets or simulation environments, fostering creativity. Reasoning mechanisms reach their highest complexity, involving generative reasoning and hypothesis exploration. Generalization becomes innovation-driven, synthesizing knowledge across fields. Applications at this stage have the highest impact, including proposing new theories, designing experiments, and driving scientific discovery (Xiong et al., 2024a; Qi et al., 2024; Pelletier et al., 2024).

3.2 Data Heterogeneity Across Four Scientific Domains

MLLMs are designed to process and integrate information from both textual and visual modalities, offering a versatile framework for handling complex scientific reasoning. However, the distinct nature of data across disciplines introduces unique challenges in model training and application. Appendix D summarizes key differences in visual features for four scientific subjects within our scope.

Each subject presents unique challenges in data representation, with mathematical tasks primarily focusing on abstract symbols and formulas, while other subjects, particularly biology, require a mix of detailed real-world imagery and conceptual explanations. These differences necessitate domain-specific adaptations in how MLLMs process and understand multimodal data (Bai et al., 2024a).

3.3 MLLM-based Scientific Reasoning

As shown in Figure 2, current MLLM-based scientific reasoning can generally be divided into the following five paradigms, which progressively enhance the reasoning capabilities of MLLMs, ultimately moving towards AGI.

Data Integration. One of the primary strengths of MLLMs is their ability to integrate multimodal data from various scientific domains. For example, in physics, models can combine textual descriptions of a problem, such as Newtons second law, with visual representations like force diagrams. The model can then reason about how different forces interact and predict motion (Barman et al., 2025; Sato, 2024). Similarly, in chemistry, MLLMs can combine chemical equations with 3D molecular structures, offering deeper insights into reaction mechanisms (Guo et al., 2023). This integrated approach allows MLLMs to handle intricate, often disjointed, data sources to generate coherent scientific explanations.

The ability to integrate and synthesize multimodal information enables the MLLM to solve complex problems more effectively. However, challenges arise when the visual data does not perfectly align with the textual explanation, potentially leading to misinterpretations (Zhang et al., 2025; Lu et al., 2023; Zhuang et al., 2024).

Knowledge Retrieval. A significant aspect of scientific reasoning is knowledge retrieval. In fields like physics and chemistry, the vast amount of scientific knowledge available - such as established theories, laws, or empirical data - can be overwhelming. Knowledge retrieval helps MLLMs access external knowledge bases, databases, and scientific literature to supplement their reasoning. For instance, when solving a chemistry problem, an MLLM could retrieve data from a chemical database to identify missing properties of substances or reactions that were not explicitly stated in the task (Prince et al., 2024; Sze and Hassoun, 2024). In addition, knowledge retrieval can aid MLLMs in bridging gaps between modalities. For example, an MLLM working on a biological problem may retrieve relevant studies from scientific papers to fill in missing knowledge, such as identifying unknown interactions between proteins or cells (Li et al., 2024j, 2025).

This aspect of MLLMs reasoning ensures the model remains up-to-date with the latest discoveries and can apply a deeper layer of scientific knowledge in reasoning processes. However, challenges in accurately selecting and integrating relevant knowledge remain, particularly when sources of conflict are present (Fan et al., 2024).

Contextual Understanding. Contextual understanding in scientific reasoning involves understanding not only the literal data presented but also

the broader context in which it is used. MLLMs are capable of this by combining visual data, such as molecular structures in chemistry, with textual descriptions of chemical properties. This allows them to reason about potential interactions between molecules in a way that goes beyond simple matching (Liu et al., 2025a; Horawalavithana et al., 2023). Wang et al. (2024b) leverage Chain-of-Thought as teaching signals to train small models to perform reasoning in complicated scenarios.

This contextual capability is crucial in fields like biology, where visual images of biological processes must be linked with underlying theories to make accurate predictions. However, this capability can be limited when the model fails to integrate textual and visual information effectively, leading to errors in reasoning (Li and Tang, 2024).

Pattern Recognition. In scientific reasoning, pattern recognition is a crucial skill that MLLMs excel at. MLLMs can detect patterns across different modalities, whether they are geometric in mathematics or experimental in chemistry. For instance, in biology, MLLMs can recognize cellular structures in images and relate them to known biological processes described in text, such as identifying mitochondria and correlating their function with energy production (Luu and Buehler, 2023; Kraus et al., 2024). Additionally, an example of pattern recognition in mathematics could involve an MLLM matching visual representations of geometric figures with algebraic equations to find solutions to geometry problems (Mouselinos et al., 2024). This capability enhances model's ability to understand complex systems across disciplines, including identifying patterns in large datasets that might be intricate for manual analysis.

This skill allows MLLMs to be highly effective in analyzing multistep reasoning problems, which are often required in scientific disciplines (Qiao et al., 2024; Yan et al., 2024b). However, pattern recognition can be hindered by noisy or low-quality visual data, particularly in domains like biology, where image clarity is critical for correct interpretation (Zhang et al., 2024e; Ren et al., 2024).

Simulation and Hypothesis Testing. MLLMs also possess the ability to perform simulation and hypothesis testing, a fundamental part of scientific reasoning (Qi et al., 2023). For example, in physics, MLLMs can simulate the effect of various forces on an object, predict outcomes, and validate those predictions against real-world data or experiments (Melko and Carrasquilla, 2024; Gao et al., 2024a; Yu et al., 2024). This capacity allows MLLMs to conduct scientific inquiries in a manner akin to human researchers, testing hypotheses and refining conclusions (Morera, 2024).

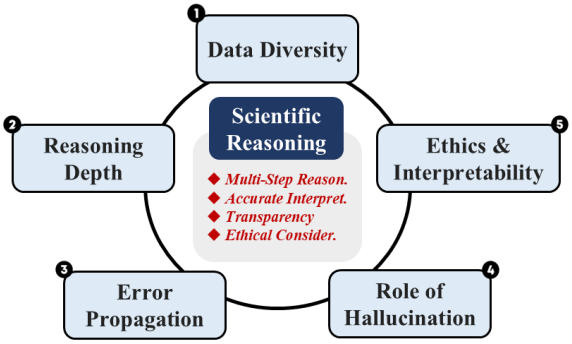

Figure 3: Challenges for MLLM-based scientific reasoning.

Despite these strengths, hypothesis testing is constrained by the quality & quantity of available data, which in some scientific domains remains insufficient or incomplete. This limits the generalizability and reliability of MLLMs in tasks requiring deep, multistep reasoning (Xiong et al., 2024a).

4 Challenges

Even though MLLMs show substantial promise in solving scientific reasoning tasks, significant challenges remain. Based on the inherent characteristics of scientific reasoning - the need for multi-step inference and precise speculation, while ensuring transparency and ethicality - we further propose the following five key challenges (Figure 3).

Data Diversity. Another challenge is the diversity of data across different scientific domains (summary of multi-domain scientific datasets can be seen in Appendix E). While mathematics is rich in textual data, such as equations and proofs, the availability of high-quality visual data is more limited (Qiao et al., 2024; Sun et al., 2024a; Liu et al., 2024c; He et al., 2024a). In contrast, fields like chemistry and biology benefit from abundant visual data, such as molecular structures and microscopic images, but the corresponding textual descriptions may not always provide the depth required for comprehensive reasoning (Alampara et al., 2024; Hocky, 2024). Without sufficient high-quality data for all modalities, model's ability to generalize across domains is compromised.

Reasoning Depth. MLLMs frequently struggle with tasks that require deep, multi-step reasoning, especially when abstract concepts are involved. In mathematics, for example, solving a theorem involves a series of logical steps that must

be followed precisely (Chen et al., 2023b). In physics, simulating complex systems, such as thermodynamics or quantum mechanics, requires a deep understanding of abstract principles and real-world conditions (Melko and Carrasquilla, 2024). MLLMs often fail to maintain this depth of reasoning, especially when applied to tasks involving intricate concepts or lengthy proof processes. This issue is particularly prevalent in fields where the complexity of reasoning extends beyond surface-level analysis and requires models to maintain rigorous logical consistency (Yang et al., 2024c). Therefore, recent work has focused on two directions to improve the length of correct reasoning chains: the development of high-quality reasoning process datasets (Yan et al., 2024b; Song et al., 2025; Zheng et al., 2024a) and the introduction of process reward models (o1 Team, 2024; Xiong et al., 2024c; Wang et al., 2024e).

Error Propagation. Error propagation is another significant challenge in multimodal reasoning. Errors in one modality, such as a misinterpreted graph or an unclear image, can propagate throughout the reasoning process, leading to incorrect conclusions (Li et al., 2024l; Yan et al., 2024b; Li et al., 2024b). For example, in a physics problem involving force vectors, an error in interpreting the vector diagram could lead to an incorrect calculation of the net force, which would then affect subsequent steps in the solution process (Jaiswal et al., 2024). The risk of error propagation is particularly high when models are tasked with handling complex, multistep problems across multiple modalities. In particular, the impact of error propagation is especially acute in fields like physics and chemistry, where the accuracy of one step can influence the entire solution process. Small errors in initial data interpretation can lead to significant discrepancies in the final outcome (Xu et al., 2024a; Li et al., 2024c).

Role of Hallucinations. One of the most complex challenges in leveraging MLLMs for scientific reasoning is determining whether hallucinations, the generation of information not grounded in the input data or knowledge base, are inherently harmful or potentially beneficial (Bai et al., 2024b; Liu et al., 2024a). While hallucinations are widely regarded as detrimental in factual tasks, their role in scientific reasoning is nuanced, particularly when considering the ultimate goal of advancing to Stage 4 of the research roadmap (*i.e.*, Creative Hypothesis Generation) (Jiang et al., 2024b). In scientific reasoning, hallucinations can undermine trust and reliability by introducing inaccuracies in critical domains. For example, in physics, a hallucinated formula or principle might lead to invalid conclusions, while in chemistry, a fabricated reaction pathway could suggest impossible or even dangerous experiments. These inaccuracies not only hinder immediate problem-solving but also propagate errors if used as a basis for further research (Li et al., 2024a; Chakraborty et al., 2024; Xu et al., 2024c). See details of hallucinations in scientific reasoning in Appendix F.

Ethical and Interpretability Issues. Ethical concerns and model interpretability are major challenges when deploying MLLMs in high-stakes scientific domains, such as medical research or chemical engineering. MLLMs often lack transparency, making it difficult for users to understand how the model arrived at a particular conclusion (AlSaad et al., 2024). Furthermore, ethical concerns arise when MLLMs are used to make decisions that could have significant consequences, such as in medical diagnoses or environmental impact assessments (Rahman et al., 2024). In biology and medicine, the potential for biased reasoning in MLLMs, especially when trained on unbalanced datasets, could lead to harmful or misleading conclusions (Stureborg et al., 2024; Wang et al., 2024d). An MLLM trained on biased medical data could fail to recognize critical symptoms in underrepresented populations, leading to erroneous diagnoses or treatment recommendations (Norori et al., 2021; Jones et al., 2024).

5 Discussion: What Next?

Building on the challenges outlined in Section 4, it is evident that while MLLMs hold great promise in advancing scientific reasoning, targeted solutions must be developed to address the limitations. Thus, we explore eight key perspectives for improving MLLMs in scientific reasoning (Figure 4).

The Necessity of Unified Scientific MLLMs. Although many domain-specific models have achieved remarkable performance in specialized scientific fields, exploring unified scientific MLLMs remains a critical pursuit (Taylor et al., 2022). Domain-specific models are optimized for particular areas (summary of scientific MLLMs can be seen in Appendix G), but they often lack the ability to integrate knowledge across disciplines (Wang et al., 2023b; Shi et al., 2024a). In contrast, a unified MLLM could facilitate inter-

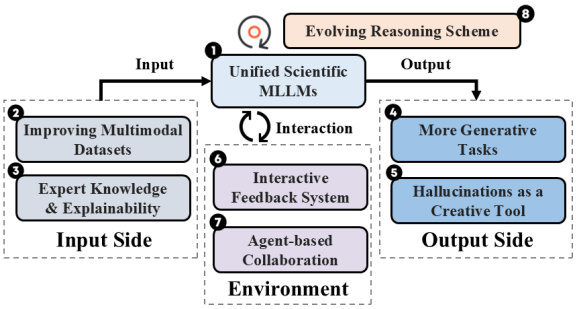

Figure 4: Eight prospects for the future of MLLMs in the field of multimodal scientific reasoning.

disciplinary reasoning, leveraging connections between fields to tackle complex problems that require holistic understanding, *e.g.,* climate change modeling (Nguyen et al., 2023) or biomedical research (Wang et al., 2023a).

For example, a unified scientific MLLM could simultaneously analyze chemical reaction pathways and their biological implications, enabling breakthroughs in drug discovery (Oniani et al., 2024; Guan and Wang, 2024). Similarly, it could integrate physics-based simulations with mathematical optimization to design more efficient renewable energy systems (Gao et al., 2024b).

Improving Multimodal Datasets. A critical step in advancing MLLMs' capabilities is the improvement of multimodal datasets (Gadre et al., 2024; Rahate et al., 2022; Bayoudh et al., 2022). Current datasets often lack the richness and variety required to train models effectively across disciplines. For instance, in chemistry, existing datasets may focus heavily on molecular structures without providing sufficient textual descriptions of reaction mechanisms (Cao et al., 2023a). Similarly, in biology, while there are abundant images of anatomical structures, these are often not paired with detailed descriptions of biological processes (Tang et al., 2023; Zhang et al., 2024g). By creating multimodal datasets that include both high-quality images and comprehensive textual descriptions across all domains, the model can better learn to correlate visual features with corresponding scientific explanations (Albalak et al., 2024; Muennighoff et al., 2023). See discussion of integration with more modalities in Appendix H.

For example, in physics, datasets that combine experimental setups with corresponding theoretical explanations could help improve a models understanding of underlying principles, such as energy conservation or force interactions. This integration would facilitate a more robust training process, ensuring that models can handle a wider range of scientific reasoning tasks. As a concrete suggestion, developing domain-specific multimodal datasets that cover not only the subject but different teaching contexts (*e.g.,* beginner vs. advanced materials) would help MLLMs generalize across varying complexity levels (Shi et al., 2023).

Integrating Expert Knowledge and Explainability. Integrating expert knowledge into MLLMs can significantly enhance their ability to reason accurately and logically (Pan et al., 2024). Expert knowledge, such as specialized theories in physics or established principles in biology, provides a framework within which the MLLM can operate. Additionally, by integrating causal reasoning into MLLMs, the models can better explain the relationships between variables, such as causes and effects, rather than simply identifying correlations (Xiong et al., 2024b; Jin et al., 2024b).

For example, in chemistry, integrating knowledge about chemical bonding, molecular dynamics, and reaction mechanisms can help guide the models predictions and interpretations (Zaki et al., 2024). Additionally, enhancing the explainability of MLLMs is crucial for transparency (Dang et al., 2024b). Ensuring that the model can justify its reasoning in a way that humans can easily understand, especially when making scientific claims, will increase its credibility and trustworthiness.

Expanding Scientific Reasoning to Generative Tasks. While significant progress has been made in using MLLMs for problem-solving, error detection, and theorem proving (Yan et al., 2024a), the potential of these models in generative tasks remains underexplored. Generative tasks such as creating curriculum-aligned questions (Mulla and Gharpure, 2023), designing comprehensive syllabi (Hu et al., 2024), or enabling digital teaching assistants (Onu et al., 2024) are highly relevant to real-world applications, *esp.* in educational contexts.

For example, MLLMs could be used to generate topic-specific exam questions that align with diverse educational standards or assist in creating adaptive learning materials tailored to different proficiency levels (Denny et al., 2024). A practical use case might involve a digital teaching assistant that generates personalized lesson plans and explanations based on students' unique learning patterns and challenges.

Hallucinations as a Creative Tool. Hallucinations may play a constructive role in fostering creativity and innovation, particularly in exploratory scientific tasks (Huang et al., 2023; Li

et al., 2024f). At the frontier of Stage 4 (shown in Section 3.1), MLLMs are expected to hypothesize beyond existing knowledge, where generating plausible but unverified information might stimulate novel ideas.

A critical challenge lies in striking the balance between mitigating harmful hallucinations and leveraging beneficial ones. This requires a fine-grained approach to model design, where hallucination mechanisms are carefully controlled based on task context. For foundational reasoning tasks, strict adherence to validated knowledge is essential, while exploratory tasks allow controlled deviations to inspire innovation (Jiang et al., 2024b).

Interactive Feedback Systems. Another powerful strategy involves the development of interactive feedback systems, which would allow MLLMs to engage with users dynamically, iterating on their answers based on user input or feedback (Abramson et al., 2022; Shtarbanov et al., 2023). This interactive feature would not only enable models to adjust their reasoning during the problem-solving process but also allow them to ask clarifying questions, improving the overall user experience and enhancing the models output.

For example, in biology, a researcher could input a biological query related to an ecological model and receive initial results from the model. This back-and-forth interaction would provide a mechanism for error correction and refinement, ensuring that outputs align more closely with expert understanding (Pan et al., 2023).

Agent-based Collaboration. A promising avenue for future development is agent-based collaboration, which involves the integration of multiple specialized agents working together to solve complex scientific problems (Guo et al., 2024b; Wang et al., 2024c). Each agent could be tailored to handle specific scientific tasks, such as mathematical reasoning, chemical reaction prediction, or biological system analysis. These agents could communicate and collaborate with each other to cross-check information, validate hypotheses, and combine knowledge from their respective domains (Chen et al., 2023a; Yan et al., 2025).

For instance, in a physics problem involving both mechanics and electromagnetism, an agent focused on classical mechanics could collaborate with an agent specialized in electromagnetism to deliver a comprehensive solution that accounts for the interactions between mechanical forces and electromagnetic fields. By building a system where multiple agents, each bringing a unique expertise, work collaboratively, MLLMs could approach complex scientific reasoning tasks with higher accuracy and robustness (Guo et al., 2024b; Zhao et al., 2024a).

Evolving Reasoning Schemes. Current reasoning architectures in MLLMs remain constrained by opaque, monolithic designs that limit adaptability to diverse scientific domains (Besta et al., 2025). Existing paradigms of Reasoning Large Models (RLMs) bifurcate into implicit RLMs (*e.g.,* QwQ (Qwen, 2024b)), where reasoning is embedded in model weights as a black box, and explicit RLMs (*e.g.,* LLaMA-Berry (Zhang et al., 2024b) & o1 (Jaech et al., 2024)), which deploy structured reasoning strategies like Monte Carlo Tree Search (MCTS) or Beam Search with modular components. While explicit methods enable stepwise evaluation, their reliance on fixed templates and proprietary training schemes hinders reproducibility and domain-specific customization.

To advance scientific reasoning, future MLLMs should integrate three innovations: (i) dynamic reasoning structures (*e.g.,* nested graphs) that adapt to multimodal inputs; (ii) process-based supervision with stepwise uncertainty metrics (*e.g.,* token-level entropy) to refine domain-specific reasoning paths; and (iii) open-source, composable toolkits for hybrid training (*i.e.,* Supervised Fine-Tuning + Reinforcement Learning phases) that decouple policy/value models, enabling collaborative, cost-efficient optimization across scientific disciplines (Besta et al., 2025).

6 Alternative Views

We also discuss two key opposing perspectives and provide counterarguments to address these concerns in Appendix I due to the space limit.

7 Conclusion

This position paper aims to emphasize the transformative potential of MLLMs in advancing scientific reasoning across diverse domains, including mathematics, physics, chemistry, and biology. Our key position is that MLLMs represent a significant step forward in enabling more comprehensive and accurate reasoning about scientific phenomena, bridging gaps between different types of data and reasoning methods. To support this stance, we reviewed the current state of MLLM applications in scientific reasoning, outlined key challenges, and proposed actionable insights.

Limitations

While this position paper presents a comprehensive vision for the role of MLLMs in scientific reasoning, we acknowledge several limitations that define the boundaries of our current analysis and highlight avenues for future exploration.

- **Focused Disciplinary Scope.** Our analysis is anchored in four core scientific disciplines: mathematics, physics, chemistry, and biology. While these fields are foundational and highly representative of multimodal reasoning challenges, they do not encompass the full spectrum of scientific inquiry. Future work could extend our proposed framework to other domains, such as earth sciences, materials science, and the social sciences, which present their own unique data modalities and reasoning paradigms.

- **Conceptual Nature of the Roadmap.** The proposed four-stage research roadmap is intended as a high-level conceptual framework to guide future development. We recognize that the progression between stages may not be strictly linear and the boundaries can be fluid. Establishing fine-grained, quantitative metrics to precisely benchmark an MLLM's position within this roadmap is a complex challenge that we leave for future research.

- **Focus on Model Capabilities over Human-AI Interaction.** Our discussion centers predominantly on the intrinsic reasoning capabilities of MLLMs. While we briefly touch upon interactive systems, a deeper exploration of the socio-technical dynamicshow these advanced models will effectively and ethically integrate into the daily workflows of human scientists and foster optimal human-AI collaborationis beyond our current scope. This represents a rich and vital avenue for future work at the intersection of AI, HCI, and the philosophy of science.

References

Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, and 1 others. 2022. Improving multimodal interactive agents with reinforcement learning from human feedback. *arXiv preprint arXiv:2211.11602*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lakshita Agarwal and Bindu Verma. 2024. From methods to datasets: A survey on image-caption generators. *Multimedia Tools and Applications*, 83(9):28077–28123.

Nawaf Alampara, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, Mara Schilling-Wilhelmi, NM Anoop Krishnan, and Kevin Maik Jablonka. Macbench: A multimodal chemistry and materials science benchmark.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Krishnan, and Kevin Maik Jablonka. 2024. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv:2411.16955*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, and 1 others. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3.

Daman Arora, Himanshu Gaurav Singh, and 1 others. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074*.

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

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2021. Molgpt: Molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, and 1 others. 2024a. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Lei Bao, Tianfan Cai, Kathy Koenig, Kai Fang, Jing Han, Jing Wang, Qing Liu, Lin Ding, Lili Cui, Ying Luo, and 1 others. 2009. Learning and scientific reasoning. *Science*, 323(5914):586–587.

Kristian G Barman, Sascha Caron, Emily Sullivan, Henk W de Regt, Roberto Ruiz de Austri, Mieke Boon, Michael Färber, Stefan Fröse, Faegheh Hasibi, Andreas Ipp, and 1 others. 2025. Large physics models: Towards a collaborative approach with large language models and foundation models. *arXiv preprint arXiv:2501.05382*.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970.

Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. 2024. Numinamath 7b tir. https://huggingface.co/AI-MO/NuminaMath-7B-TIR.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, and 1 others. 2025. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. Xai meets llms: A survey of the relation between explainable ai and large language models. *arXiv preprint arXiv:2407.15248*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023a. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023b. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *ArXiv*, abs/2311.16208.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2024. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *arXiv preprint arXiv:2403.16527*.

Jinho Chang and Jong-Chul Ye. 2022. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2023a. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

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

Yong Chen, Hongpeng Chen, and Songzhi Su. 2023c. Fine-tuning large language models in education. In *2023 13th International Conference on Information Technology in Medicine and Education (ITME)*, pages 718–723. IEEE.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2024c. Scienceagentbench: Toward rigorous assessment of language

agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

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.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Christian Norgaard, Nayantara Mudur, Martyna Beata Plomecka, Paul Raccuglia, and 1 others. 2025. Curie: Evaluating llms on multitask scientific long-context understanding and reasoning. In *The Thirteenth International Conference on Learning Representations*.

Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Aiwei Liu, and Xuming Hu. 2024a. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. *arXiv preprint arXiv:2411.02708*.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, and 1 others. 2024b. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Nour El Houda Dehimi and Zakaria Tolba. 2024. Attention mechanisms in deep learning: Towards explainable artificial intelligence. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–7. IEEE.

Paul Denny, Stephen MacNeil, Jaromir Savelka, Leo Porter, and Andrew Luxton-Reilly. 2024. Desirable characteristics for ai teaching assistants in programming education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 408–414.

Peng Di, Jianguo Li, Hang Yu, Wei Jiang, Wenting Cai, Yang Cao, Chaoyu Chen, Dajun Chen, Hongwei Chen, Liang Chen, and 1 others. 2024. Codefuse-13b: A pretrained multi-lingual code large language model. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 418–429.

Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. 2025. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *Preprint*, arXiv:2505.23579.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. 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.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, and 1 others. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from agi. *arXiv preprint arXiv:2405.10313*.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Jia-Yi Fu, Lei Lin, Xiaoyang Gao, Pengli Liu, Zhengzong Chen, Zhirui Yang, Shengnan Zhang, Xue Zheng, Yan Li, Yuliang Liu, Xucheng Ye, Yiqiao Liao, Chao Liao, Bin Chen, Chengru Song, Junchen Wan, Zijia Lin, Fuzheng Zhang, Zhongyuan Wang, and 2 others. 2023. Kwaiyiimath: Technical report. *ArXiv*, abs/2310.07488.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024a. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024b. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE.

Steven N Goodman. 2016. Aligning statistical and scientific reasoning. *Science*, 352(6290):1180–1181.

Shenghui Guan and Guanyu Wang. 2024. Drug discovery and development in the era of artificial intelligence: From machine learning to large language models. *Artificial Intelligence Chemistry*, 2(1):100070.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, and 1 others. 2024a. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

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*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. 2025a. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. *ArXiv*, abs/2503.10627.

Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Hongsheng Li, and Pheng-Ann Heng. 2025b. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. *arXiv preprint*.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2024. Urbanvlp: A multi-granularity vision-language pretrained foundation model for urban indicator prediction. *arXiv preprint arXiv:2403.16831*.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024a. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, and 1 others. 2024b. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024c. Cmmu: A benchmark for chinese multimodal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Glen M Hocky. 2024. Connecting molecular properties with plain language. *Nature Machine Intelligence*, 6(3):249–250.

Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. 2023. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.

Shuaibo Hu and Kui Yu. 2024. Learning robust rationales for model explainability: A guidance-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18243–18251.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Yinkui Huang, Tianrun Gao, Jiangjiang Zhang, Xiaohong Liu, and Guangyu Wang. 2024a. Adapting large language models for biomedicine though retrieval-augmented generation with documents scoring. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5770–5775. IEEE.

Zhen Huang, Zengzhi Wang, Shijie Xia, and Pengfei Liu. 2024b. Olympicarena medal ranks: Who is the most intelligent ai so far? *arXiv preprint arXiv:2406.16772*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, page 102270.

Aryan Jadon, Avinash Patil, and Shashank Kumar. 2025. Enhancing domain-specific retrieval-augmented generation: Synthetic data generation and evaluation using reasoning models. *arXiv preprint arXiv:2502.15854*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. 2024. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*.

Boran Jiang, Yuqi Wang, Yi Luo, Dawei He, Peng Cheng, and Liangcai Gao. 2024a. Reasoning on efficient knowledge paths: knowledge graph guides large language model for domain question answering. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 142–149. IEEE.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024b. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Shiyu Jin, Jinxuan Xu, Yutian Lei, and Liangjun Zhang. 2024a. Reasoning grasping via multimodal large language model. *arXiv preprint arXiv:2402.06798*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2024b. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36.

Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, and 1 others. 2024. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. 2022. Hypertree proof search for neural theorem proving. *ArXiv*, abs/2205.11491.

Antone E Lawson. 2004. The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education*, 2:307–338.

Martha Lewis and Melanie Mitchell. 2024. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024a. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7668–7681.

Hang Li, Tianlong Xu, Kaiqi Yang, Yucheng Chu, Yanling Chen, Yichi Song, Qingsong Wen, and Hui Liu. 2024b. Ask-before-detection: Identifying and mitigating conformity bias in llm-powered error detector for math word problem solutions. *arXiv preprint arXiv:2412.16838*.

Hang Li, Tianlong Xu, Chaoli Zhang, Eason Chen, Jing Liang, Xing Fan, Haoyang Li, Jiliang Tang, and Qingsong Wen. 2024c. Bringing generative ai to adaptive learning in education. *arXiv preprint arXiv:2402.14601*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024d. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024e. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024f. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, and Jing Ma. 2024g. Mmcode: Evaluating multi-modal code large language models with visually rich programming problems. *arXiv preprint arXiv:2404.09486*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024h. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024i. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, page 104769.

Mingchen Li, Zaifu Zhan, Han Yang, Yongkang Xiao, Jiatan Huang, and Rui Zhang. 2024j. Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness. *arXiv preprint arXiv:2405.08151*.

Sihang Li, Jian Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024k. Scilitllm: How to adapt llms for scientific literature understanding. *ArXiv*, abs/2408.15545.

Songtao Li and Hao Tang. 2024. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

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

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024m. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024a. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024b. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

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

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Hongxuan Liu, Haoyu Yin, Zhiyao Luo, and Xiaonan Wang. 2025a. Integrating chemistry knowledge in large language models via prompt engineering. *Synthetic and Systems Biotechnology*, 10(1):23–38.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, and 1 others. 2025b. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Peng Liu, Yiming Ren, and Zhixiang Ren. 2023c. Gitmol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073.

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*.

Xiang Liu, Penglei Sun, Shuyan Chen, Longhan Zhang, Peijie Dong, Huajie You, Yongqi Zhang, Chang Yan, Xiaowen Chu, and Tong yi Zhang. 2025c. Perovskite-llm: Knowledge-enhanced large language models for perovskite solar cell research. *Preprint*, arXiv:2502.12669.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025d. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Micha Livne, Zulfat Miftahutdinov, E. Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony B Costa, Alex Aliper, and Alex Zhavoronkov. 2023. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15:8380 – 8389.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

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 math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

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*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang 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*, abs/2308.09583.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*.

Rachel K. Luu and M. Buehler. 2023. Bioinspiredllm: Conversational large language model for the mechanics of biological and bioinspired materials. *Advanced Science*, 11.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein language model for multi-task protein language processing. *ArXiv*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. *Preprint*, arXiv:2505.14652.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, and 1 others. 2024. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*.

Roger G Melko and Juan Carrasquilla. 2024. Language models for quantum simulation. *Nature Computational Science*, 4(1):11–18.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Santiago Miret and Nandan M Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, and 1 others. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

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*.

Mistral AI Team. 2024. Mathσtral.

MoonshotAI. 2024. k0-math official platform.

Albert Morera. 2024. Foundation models in shaping the future of ecology. *Ecological Informatics*, page 102545.

Meredith Ringel Morris. 2023. Scientists' perspectives on the potential for generative ai in their fields. *arXiv preprint arXiv:2304.01420*.

Spyridon Mouselinos, Henryk Michalewski, and Mateusz Malinowski. 2024. Beyond lines and circles: Unveiling the geometric reasoning gap in large language models. *arXiv preprint arXiv:2402.03877*.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *arXiv preprint arXiv:2501.03151*.

Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G. Rodriques, and Andrew D. White. 2025. Training a scientific reasoning model for chemistry. *Preprint*, arXiv:2506.17238.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. 2023. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*.

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*, abs/2405.20335.

Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. 2024. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *arXiv preprint arXiv:2405.06545*.

Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. 2021. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2(10).

Skywork o1 Team. 2024. Skywork-o1 open series. https://huggingface.co/Skywork.

David Oniani, Jordan Hilsman, Chengxi Zang, Junmei Wang, Lianjin Cai, Jan Zawala, and Yanshan Wang. 2024. Emerging opportunities of using large language language models for translation between drug molecules and indications. *arXiv preprint arXiv:2402.09588*.

Peter Onu, Anup Pradhan, and Charles Mbohwa. 2024. Potential to use metaverse for future teaching and learning. *Education and Information Technologies*, 29(7):8893–8924.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. 2024. Survey of large multimodal model datasets, application categories and taxonomy. *arXiv preprint arXiv:2412.17759*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. 2024. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*.

Alexander R Pelletier, Joseph Ramirez, Irsyad Adam, Simha Sankar, Yu Yan, Ding Wang, Dylan Steinecke, Wei Wang, and Peipei Ping. 2024. Explainable biomedical hypothesis generation via retrieval augmented generation enabled large language models. *arXiv preprint arXiv:2407.12888*.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *ArXiv*, abs/2009.03393.

Vignesh Prabhakar, Md Amirul Islam, Adam Atanas, Yao-Ting Wang, Joah Han, Aastha Jhunjhunwala, Rucha Apte, Robert Clark, Kang Xu, Zihan Wang, and Kai Liu. 2025. Omniscience: A domain-specialized llm for scientific reasoning and discovery. *Preprint*, arXiv:2503.17604.

Michael H Prince, Henry Chan, Aikaterini Vriza, Tao Zhou, Varuni K Sastry, Yanqi Luo, Matthew T Dearing, Ross J Harder, Rama K Vasudevan, and

Mathew J Cherukara. 2024. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Computational Materials*, 10(1):251.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: a comprehensive evaluation. *arXiv preprint arXiv:2407.08940*.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Jimin Huang, Qianqian Xie, and Jianyun Nie. 2025. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv preprint arXiv:2502.08127*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Qwen. 2024a. Qvq: To see the world with wisdom.

Qwen. 2024b. Qwq: Reflect deeply on the boundaries of the unknown, november 2024. *URL https://qwenlm. github. io/blog/qwq-32b-preview*.

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239.

Md Abdur Rahman, Lamyaa Alqahtani, Amna Albooq, and Alaa Ainousah. 2024. A survey on security and privacy of large multimodal deep learning models: Teaching and learning perspective. In *2024 21st Learning and Technology Conference (L&T)*, pages 13–18. IEEE.

Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science*.

Chandan K Reddy and Parshin Shojaee. 2024. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. *arXiv preprint arXiv:2412.11427*.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383.

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2024. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–563.

Nikolaos Rodis, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, and Georgios Th Papadopoulos. 2024. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. *IEEE Access*.

Keisuke Sato. 2024. Exploring the educational landscape of ai: Large language models' approaches to explaining conservation of momentum in physics. *arXiv preprint arXiv:2407.05308*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024a. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Li Bing, and Roy Ka wei Lee. 2024b. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. 2024. Llm-sr: Scientific equation discovery via programming with large language models. *ArXiv*, abs/2404.18400.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. 2025. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*.

Ali Shtarbanov, Mengjia Zhu, Nicholas Colonnese, and Amirhossein Hajiagha Memar. 2023. Sleeveio: modular and reconfigurable platform for multimodal wearable haptic feedback interactions. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*, pages 1–15.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. 2024. A large encoder-decoder family of foundation models for chemical language. *ArXiv*, abs/2407.20267.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *ArXiv*, abs/2405.01724.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2023. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*.

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.

Ash Sze and Soha Hassoun. 2024. Evaluation of search-enabled pre-trained large language models on retrieval tasks for the pubchem database. *bioRxiv*, pages 2024–08.

TALEducation. 2023. Mathgpt official platform.

Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, and 1 others. 2023. Explainable multi-task learning for multimodality biological data analysis. *Nature communications*, 14(1):2546.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171.

J Vaishnavi and V Narmatha. 2024. Video captioning–a survey. *Multimedia Tools and Applications*, pages 1–32.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, and 1 others. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Feng Wang. 2024. Lighthouse: A survey of agi hallucination. *arXiv preprint arXiv:2401.06792*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024a. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

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

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024c. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024d. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024e. 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024f. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024g. 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. 2024h. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jingxuan Wei, Cheng Tan, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z Li. 2024. Enhancing human-like multimodal reasoning: a new challenging dataset and comprehensive framework. *Neural Computing and Applications*, 36(33):20849–20861.

Andrew D White. 2023. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458.

Robin H Willemsen, Isabelle C de Vink, Evelyn H Kroesbergen, and Ard W Lazonder. 2023. The role of creative thinking in children's scientific reasoning. *Thinking Skills and Creativity*, 49:101375.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and 1 others. 2024a. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*.

Yifan Wu, Min Zeng, Yang Li, Yang Zhang, and Min Li. 2024b. Molmetalm: a physicochemical knowledge-guided molecular meta language model. *ArXiv*, abs/2411.15500.

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*, abs/2410.15700.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, Imran Razzak, and Bram Hoex. 2023. Darwin series: Domain specific large language models for natural science. *ArXiv*, abs/2308.13565.

Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Wenjie Zhang, Chunyu Kit, Dongzhan Zhou, and Bram Hoex. 2024. Darwin 1.5: Large language models as materials science adapted learners. *ArXiv*.

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

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. 2024b. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv*.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024a. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*.

Siheng Xiong, Delin Chen, Qingyang Wu, Longxuan Yu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2024b. Improving causal reasoning in large language models: A survey. *arXiv preprint arXiv:2410.16676*.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. 2024c. An implementation of generative prm. https://github.com/RLHFlow/RLHF-Reward-Modeling.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

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

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. 2024b. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *ArXiv*, abs/2404.02893.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024c. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

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, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, and 1 others. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. 2025. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. 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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *ArXiv*, abs/2409.12122.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. 2024c. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.

Liu Yang, Haihua Yang, Wenjun Cheng, Lei Lin, Chenxia Li, Yifu Chen, Lunan Liu, Jianfei Pan, Tianwen Wei, Biye Li, Liang Zhao, Lijie Wang, Bo Zhu, Guoliang Li, Xuejie Wu, Xilin Luo, and Rui Hu. 2023a. Skymath: Technical report. *arXiv preprint arXiv: 2310.16713*.

Yifei Yang, Runhan Shi, Z. Li, Shu Jiang, Bao-Liang Lu, Yang Yang, and Hai Zhao. 2024d. Batgpt-chem: A foundation large model for retrosynthesis prediction. *ArXiv*, abs/2408.10285.

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

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024f. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *ArXiv*, abs/2410.07076.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Robert Tang. 2025. MMSciBench: Benchmarking language models on Chinese multimodal scientific problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14621–14663, Vienna, Austria. Association for Computational Linguistics.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, and 3 others. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *ArXiv*, abs/2402.06332.

Long Long Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zheng Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *ArXiv*, abs/2309.12284.

Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C Will, Gunnar Behrens, Julius Busecke, and 1 others. 2024. Climsim: A large multi-scale dataset for hybrid physics-ml climate emulation. *Advances in Neural Information Processing Systems*, 36.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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*, abs/2309.05653.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 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*, abs/2405.03548.

Mohd Zaki, NM Anoop Krishnan, and 1 others. 2024. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, Mao Su, Han sen Zhong, Yuqiang Li, and Wanli Ouyang. 2024a. Chemllm: A chemical large language model. *ArXiv*, abs/2402.06852.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, and 1 others. 2024b. Llamaberry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024c. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and 1 others. 2024d. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and 1 others. 2024e. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024f. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Yuchen Zhang, Ratish Kumar Chandrakant Jha, Soumya Bharadwaj, Vatsal Sanjaykumar Thakkar, Adrienne Hoarfrost, and Jin Sun. 2024g. A benchmark dataset for multimodal prediction of enzymatic function coupling dna sequences and natural language. *arXiv preprint arXiv:2407.15888*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024a. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

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. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wayne Xin Zhao, Kun Zhou, Beichen Zhang, Zheng Gong, Zhipeng Chen, Yuanhang Zhou, Ji rong Wen, Jing Sha, Shijin Wang, Cong Liu, and Guoping Hu. 2023b. Jiuzhang 2.0: A unified chinese pre-trained language model for multi-task mathematical problem solving. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. *Preprint*, arXiv:2501.12380.

Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Ziping Wan, Yansi Li, Zhongyang Dai, Xin Chen, and Kai Yu. 2024b. Chemdfm-x: Towards large multimodal model for chemistry. *ArXiv*, abs/2409.13194.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024c. Chemdfm: A large language foundation model for chemistry. *ArXiv*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024a. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024b. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*.

Siru Zhong, Xixuan Hao, Yibo Yan, Ying Zhang, Yangqiu Song, and Yuxuan Liang. 2024a. Urbancross: Enhancing satellite image-text retrieval with cross-domain adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6307–6315.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024b. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024a. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

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

Xibin Zhou, Chenchen Han, Yingqi Zhang, Jin Su, Kai Zhuang, Shiyu Jiang, Zichen Yuan, Wei Zheng, Fengyuan Dai, Yuyang Zhou, Yuyang Tao, Dan Wu, and Fajie Yuan. 2025. Decoding the molecular language of proteins with evolla. *bioRxiv*.

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*.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, and 1 others. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606.

Contents of Technical Appendices

A Clarification of the Position Scope **24**
- A.1 Novelty of Our Position 24
- A.2 Unique Challenges for MLLMs in Scientific Reasoning vs. General Challenges 25
- A.3 Significance and Examples of Reasoning across Scientific Domains . 26
- A.4 Integration with Experimentation with Our Roadmap 26
- A.5 Quantitative Analysis of Multimodal Components 26

B Formulation of Scientific Reasoning Task **27**

C Four Phases of Research Roadmap **28**
- C.1 Phase 1: Broad Knowledge and Recognition 28
- C.2 Phase 2: Analogical Reasoning and Generalization 28
- C.3 Phase 3: Insightful Inference . . . 28
- C.4 Phase 4: Creative Hypothesis Generation 29
- C.5 Summary 29

D Data Differences among Four Domains **29**
- D.1 Mathematics: Structured Abstraction 29
- D.2 Physics: Real-world Dynamics . . 29
- D.3 Chemistry: Molecular and Symbolic 29
- D.4 Biology: Conceptual Complexity . 29
- D.5 Summary and Future Directions . 29

E Multi-Domain Scientific Reasoning Benchmark **30**

F Details of Hallucinations in Scientific Reasoning **30**
- F.1 Different Types of Hallucinations in Scientific Reasoning 30
- F.2 Existing Mitigation Strategies . . 32
- F.3 Why Current Mitigation Methods Fall Short in Scientific Reasoning 32

G Multimodal Scientific (M)LLMs **33**

H Discussion of More Modalities in Scientific Reasoning **33**
- H.1 Current Limitations and the Imperative for Richer Modalities 33
- H.2 Technical Approaches for Broader Modality Integration 34
- H.3 Advanced Cross-Modal Interaction and Scientific Reasoning . . . 34
- H.4 Task-Driven Training Strategies and Future Directions 35

I Discussion of Alternative Views **35**
- I.1 Domain-Specific Models as a Superior Alternative 35
- I.2 Risks of Over-reliance on MLLMs 37

J Clarification of LLM Usage **38**

Technical Appendices and Supplements

In this appendix, we first provide a clarification of our position's scope (Appendix A), detailing its novelty, the unique challenges MLLMs face in scientific reasoning compared to general tasks, the significance of cross-domain reasoning, and the integration of experimentation with our proposed roadmap. Then, in Appendix B, we offer a formal formulation of the scientific reasoning task, while Appendix C elaborates on our four-stage research roadmap. Appendix D discusses data heterogeneity across key scientific domains, and Appendix E reviews relevant multi-domain scientific reasoning benchmarks, with Appendix G providing an overview of current scientific (M)LLMs. In Appendix F, we present an in-depth analysis of hallucinations in scientific reasoning, including their types, existing mitigation strategies, and why they often fall short in scientific contexts. Appendix H further explores the necessity and technical approaches for integrating richer modalities such as audio, video, and 3D data. Finally, Appendix I addresses alternative views and counterarguments, such as the preference for hyper-specialized models and concerns regarding MLLM reliability and explainability.

A Clarification of the Position Scope

A.1 Novelty of Our Position

This position paper presents a novel and compelling argument for the transformative potential of MLLMs in advancing the full spectrum of scientific reasoning, coupled with a unique, MLLM-centric four-stage roadmap towards this goal. While individual concepts such as applying AI to science or outlining stages of AI development have been discussed elsewhere, our work distinguishes itself through a specific synthesis, focus, and scope that addresses critical gaps in the current discourse:

1. **Primacy of Multimodality for Comprehensive Scientific Reasoning:** Our central thesis is that the inherent ability of MLLMs to integrate and reason over diverse data modalities is paramount for a holistic advancement in scientific reasoning. This contrasts sharply with existing literature that often focuses on unimodal approaches or narrower applications.
    - For instance, comprehensive surveys on "Scientific LLMs" (Zhang et al., 2024f) predominantly analyze text-based models and their applications in scientific discovery, largely overlooking the rich multimodal nature of scientific data and reasoning. Our position explicitly champions the multimodal perspective as indispensable.
    - Similarly, while works like (Reddy and Shojaee, 2024) discuss generative AI for "scientific discovery," their scope is primarily confined to this (albeit important) sub-task. Our position encompasses a broader range of scientific reasoning capabilities crucial for overall scientific advancement, including foundational knowledge acquisition, commonsense scientific question answering, analogical reasoning across disciplines, and insightful inference from complex data all areas where MLLMs offer unique advantages.
2. **A Dedicated Four-Stage Roadmap for MLLM-Driven Scientific Reasoning Towards AGI:** We propose a structured four-stage research roadmap (Broad Knowledge and Recognition, Analogical Reasoning and Generalization, Insightful Inference, and Creative Hypothesis Generation) that is specifically tailored to the evolving capabilities of MLLMs within *all scientific reasoning scenarios*, ultimately aiming towards Artificial General Intelligence (AGI). This roadmap is distinct from others in its MLLM-centricity and breadth:
    - It moves beyond general discussions on generative AI's potential in science (Morris, 2023), which, while insightful, did not offer a structured, MLLM-focused developmental pathway.
    - It differs significantly from task-specific pipelines. For example, the highly valuable "AI Scientist" framework proposed by (Lu et al., 2024) outlines a three-stage pipeline (idea generation $\rightarrow$ experiment iteration $\rightarrow$ paper write-up) primarily designed for *automated scientific discovery and specific generative outputs* like code and papers. Our roadmap is more encompassing: it situates such advanced discovery and generation capabilities (which we discuss in Section 5, including applications like digital teaching assistants)

within its sophisticated later stages (particularly Stage 4). Crucially, our roadmap also articulates the foundational and intermediate reasoning abilities (Stages 1-3) that are prerequisite for achieving such creative and autonomous scientific exploration with MLLMs. These earlier stages, focusing on multimodal data integration, knowledge retrieval, contextual understanding, and analogical generalization, are essential for building robust and reliable scientific reasoning systems.

In essence, the novelty of our position lies not in addressing these themes in isolation, but in the **holistic and integrated argument** that MLLMs, by virtue of their multimodal capabilities, are uniquely positioned to revolutionize scientific reasoning across its entire breadth. Furthermore, we provide a **dedicated, MLLM-specific, and progressive developmental trajectory** towards realizing this vision. This distinct stance, we believe, fills a crucial gap and warrants greater exposure and discussion within the machine learning community, aligning perfectly with the objectives of a Position Paper to highlight compelling arguments that can shape future research directions.

A.2 Unique Challenges for MLLMs in Scientific Reasoning vs. General Challenges

While MLLMs face a spectrum of general challenges, their application to the domain of scientific reasoning introduces distinct and often amplified difficulties. We identify four key areas where the challenges for MLLMs in scientific reasoning diverge significantly from those encountered in more general applications:

1. **Demand for Rigor, Precision, and Verifiability:** In general tasks such as summarization or creative writing, MLLMs may employ approximate reasoning and provide plausible-sounding answers where occasional logical flaws or imprecision might be acceptable. However, scientific reasoning mandates an exceptionally high standard. It requires strict logical consistency, mathematical precision, and the ability to produce verifiable step-by-step derivations. The outputs must not only be correct but also demonstrably so, adhering to the rigorous validation principles of the scientific method.

2. **Highly Specialized and Structured Multimodal Data:** General-purpose MLLMs are typically designed to handle diverse, often unstructured or semi-structured, multimodal data, with alignment efforts often focusing on broad semantic correspondence. In contrast, scientific reasoning necessitates the processing and deep understanding of highly structured, symbolic, and specialized data modalities. This includes, but is not limited to, complex chemical formulas, genetic sequences, intricate diagrams with specific notational conventions, and precise experimental data. Effective integration and reasoning over such specialized data require more than general semantic understanding; they demand the ability to parse and interpret domain-specific syntax and semantics.

3. **The Critical Role of Causality and Mechanistic Understanding:** While MLLMs often excel at pattern recognition and identifying correlations within large datasets—valuable capabilities in many contexts—scientific reasoning demands a more profound level of understanding. A core objective of science is to move beyond mere correlation to infer causality and elucidate the underlying mechanisms that govern phenomena. MLLMs applied to science must therefore develop capabilities to not just describe *what* happens, but to reason about *why* and *how* it happens, forming a crucial distinction from general pattern-matching tasks.

4. **Nuanced Role and High Stakes of Hallucination:** Hallucinations, or the generation of factually incorrect or nonsensical information, are generally undesirable in any MLLM application as they lead to misinformation. However, in the context of scientific reasoning, factual hallucinations are particularly detrimental due to the high stakes involved; inaccuracies can misdirect research, lead to flawed conclusions, and erode trust. Paradoxically, there is also a unique potential role for *controlled* generative capabilities—which might share some characteristics with hallucination if unconstrained—in the creative hypothesis generation phase. The challenge lies in fostering this creative, "out-of-the-box" thinking while strictly preventing uncontrolled factual inaccuracies, demanding a sophisticated balance not typically required in general applications.

A.3 Significance and Examples of Reasoning across Scientific Domains

The capacity to synthesize knowledge and reason across disparate scientific domains is fundamental to addressing complex, multifaceted global challenges and accelerating the pace of discovery. Many scientific frontiers lie at the intersection of disciplines, requiring an integrated understanding of phenomena that transcend traditional boundaries. MLLMs are uniquely positioned to facilitate this interdisciplinary synergy by their inherent ability to process, correlate, and reason over diverse data typestextual, visual, numerical, and symbolicfrom various fields. This capability can unlock novel insights and solutions that might remain obscured when viewed through a single disciplinary lens. Below, we illustrate this potential with representative examples:

1. **Drug Discovery and Precision Medicine:** MLLMs can integrate visual molecular structures, textual biomedical literature on biological pathways, and numerical patient genomic data. This allows for accelerated identification of promising drug candidates and prediction of patient-specific responses, paving the way for personalized medicine.

2. **Materials Science and Engineering:** By analyzing material microstructure images (visual), crystal structure information (symbolic/visual), and textual data from scientific literature on synthesis and properties, MLLMs can rapidly predict characteristics of novel materials. This capability can guide experimental design and accelerate the discovery of materials with desired functionalities.

3. **Climate Modeling and Earth System Science:** MLLMs can fuse satellite imagery (visual), numerical weather pattern data, and textual reports on atmospheric composition. This synthesis enables the construction of more comprehensive climate models, uncovering complex correlations and improving predictions of climate change impacts and extreme weather events.

A.4 Integration with Experimentation with Our Roadmap

Experimentation is undeniably central to scientific advancement. While our four-stage roadmap (Broad Knowledge and Recognition, Analogical Reasoning, Insightful Inference, Creative Hypothesis Generation) focuses on the evolution of MLLM *reasoning capabilities*, it inherently supports the experimental process, which we view as a sophisticated workflow leveraging these abilities.

For instance, designing novel experiments draws upon Creative Hypothesis Generation (Stage 4) and Insightful Inference (Stage 3), while interpreting complex experimental data relies heavily on Insightful Inference. Our framework explicitly addresses this in Section 3.3 ("Simulation & Hypothesis Testing"), where MLLMs contribute to generating testable hypotheses, designing *in silico* experiments, and predicting outcomes—tasks demanding advanced inference and creative generation (Stages 3 & 4). The progression through our stages equips MLLMs to assist in experimental design (Stages 1 & 2), facilitate simulations (Stages 3 & 4), enhance multimodal data analysis (Stage 3), and support iterative refinement of hypotheses and experiments (Stages 3 & 4). Furthermore, our Discussion (Section 5) on the future outlook of MLLMs (input, model, output, environment) naturally aligns with the components of scientific experimentation, where MLLMs process experimental inputs, perform reasoning, generate outputs like protocols or analyzed results, and operate within the scientific environment.

In essence, the capacity to support and augment scientific experimentation is an emergent outcome as MLLMs advance through the foundational reasoning capabilities outlined in our roadmap, making them integral to the experimental lifecycle.

A.5 Quantitative Analysis of Multimodal Components

To substantiate our position on the challenges and opportunities for MLLMs in scientific reasoning, this section provides a quantitative analysis of how MLLMs process and rely on different modalities, with a particular focus on the visual component. The analysis draws upon empirical experiment from two comprehensive benchmarks, **SciVerse** (Guo et al., 2025b) and **MathVerse** (Zhang et al., 2025), which systematically deconstruct problem formulations to isolate and evaluate the models' visual reasoning capabilities.

The central hypothesis investigated is that while MLLMs are designed to be multimodal, their per-

formance in complex scientific and mathematical domains is disproportionately reliant on textual information, often failing to genuinely "see" or reason from diagrams. This creates a significant performance gap when crucial information is shifted from text to the visual modality, mirroring many real-world scenarios.

Analysis from SciVerse: General Scientific Reasoning. The SciVerse benchmark evaluates MLLMs on scientific problems (Physics, Chemistry, Biology) by creating versions with varying degrees of visual dependency. The `Knowledge-free` version presents a standard problem with textual descriptions and a diagram. The `Vision-rich` version moves some conditional information from the text into the diagram, and the `Vision-only` version embeds the entire question within the visual input. As shown in Table 1, the performance of even top-tier models like GPT-4o degrades as the reliance on visual perception increases.

The data reveals a consistent trend: when MLLMs are forced to extract information from diagrams rather than text, their accuracy declines. GPT-4o's accuracy drops by 5.0 percentage points, indicating that even state-of-the-art models struggle to interpret and integrate visually presented scientific conditions. This suggests that the visual encoder and cross-modal fusion mechanisms in current MLLMs are a significant bottleneck for general scientific reasoning.

Analysis from MathVerse: Mathematical Reasoning. The MathVerse benchmark provides an even more granular analysis for mathematical problems, a domain where diagrams are often structured and indispensable. It progressively shifts information from text to vision across six versions. For this analysis, we focus on the versions that best illustrate the visual dependency: `Text-dominant` (full text), `Text-lite` (redundant text removed), `Vision-intensive` (implicit properties in vision), `Vision-dominant` (key values in vision), and `Vision-only` (all in vision).

The results, as detailed in Table 2 and Figure 5, are striking. The steady and significant performance degradation as we move from `Text-dominant` to `Vision-only` provides compelling evidence that MLLMs heavily rely on textual "shortcuts." GPT-4V's accuracy plummets by over 23 percentage points, demonstrating a critical failure to parse essential conditions, properties, and questions from mathematical diagrams.

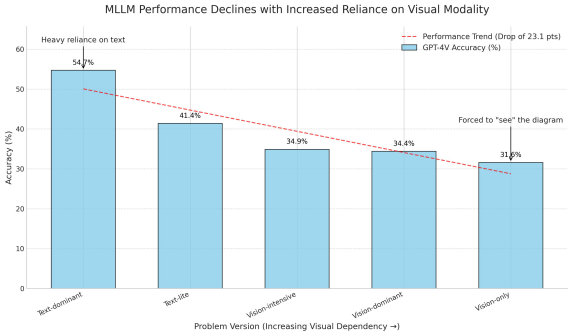

Figure 5: MLLM performance declines with increased reliance on visual modality from MathVerse benchmark.

MathVerse further reveals that some models, such as Qwen-VL-Max, astonishingly perform better without the diagram input at all (+5.1%), treating the visual modality as a source of distraction rather than information.

The quantitative data from both SciVerse and MathVerse converges on a clear conclusion: a critical "vision-language gap" exists in the scientific reasoning capabilities of current MLLMs. Their strength lies predominantly in language processing, while their ability to perform deep, structured reasoning based on visual inputsespecially the information-rich diagrams common in science and mathremains underdeveloped. This quantitative analysis underscores the challenges highlighted in our position paper and reinforces the need for future research to focus on enhancing genuine multimodal understanding, moving beyond mere text-based reasoning with visual adornments.

B Formulation of Scientific Reasoning Task

Mathematically, the scientific reasoning process can be modeled as an optimization task over a multimodal knowledge graph $G = (V, E)$, where V denotes nodes (concepts, entities, or data points) and E represents edges (relationships or interactions). Let X_m represent the modality-specific input data, including mandatory modalities such as textual descriptions X_t and visual representations X_v, as well as optional modalities like numerical data X_n or other specialized inputs X_o. The goal is to predict or infer target outputs Y based on a reasoning function f_θ, parameterized by θ, over G:

$$Y = f_\theta(G, X_m) = f_\theta(V, E, X_t, X_v, X_n, X_o).$$

Mathematics: Consider solving a geometry problem where X_t provides the problem description, X_v includes the geometric diagram, and X_n optionally contains measurements or coordinates.

Table 1: Accuracy of MLLMs on SciVerse versions with increasing visual dependency. The performance drop from `Knowledge-free` to `Vision-only` highlights the challenge of visual information processing and reasoning.

Model	Knowledge-free (Acc %)	Vision-rich (Acc %)	Vision-only (Acc %)	Perf. Drop (KF to VO)
GPT-4o	55.2	54.4	50.2	**-5.0 pts**
LLaVA-OneVision (7B)	47.2	46.5	41.9	**-5.3 pts**

Table 2: Accuracy of MLLMs across MathVerse versions, showing a progressive decline as problems become more vision-reliant.

Model	Text-dom (%)	Text-lite (%)	Vision-int (%)	Vision-dom (%)	Vision-only (%)	Overall Drop
GPT-4V	54.7	41.4	34.9	34.4	31.6	**-23.1 pts**
InternLM-XC2 (7B)	22.3	17.0	15.7	16.4	11.0	**-11.3 pts**

The reasoning function f_θ integrates these inputs to infer the solution, such as identifying the area of a triangle.

Physics: In a physics experiment, X_t describes the theoretical background, X_v presents the experimental setup image, and X_n includes sensor data such as velocity or temperature measurements. The function f_θ predicts the outcome or validates a hypothesis.

Chemistry: A multimodal analysis of a chemical reaction could include X_t for the reaction mechanism, X_v for molecular structure visualizations, and X_o for spectroscopic data. The model predicts reaction yield or product properties.

Biology: When studying gene expression, X_t describes the biological context, X_v contains microscope images, and X_n optionally includes numerical gene expression levels. The reasoning function predicts gene interactions or cellular behavior.

These examples illustrate how MLLMs can handle diverse inputs, integrating mandatory and optional modalities to perform complex scientific reasoning tasks.

C Four Phases of Research Roadmap

MLLMs have seen significant advancements in recent years, positioning their reasoning capabilities as a pivotal element on the pathway to achieving Artificial General Intelligence (AGI) (Wang et al., 2024h; Yan et al., 2024a; Sun et al., 2023; Jin et al., 2024a; Yan et al., 2024c; Wei et al., 2024). However, realizing AGI requires navigating a structured roadmap characterized by progressively complex reasoning tasks. This section outlines the Four Phases of Research Roadmap, each delineated by unique data requirements, reasoning mechanisms, generalization abilities, and impact. Table 3 provides a comparative overview of these phases.

C.1 Phase 1: Broad Knowledge and Recognition

In this phase, MLLMs prioritize high diversity and low specificity in data and knowledge. Tasks involve integrating vast and diverse datasets, emphasizing retrieval and alignment mechanisms. For example, MLLMs excel in synthesizing encyclopedic knowledge, aligning visual inputs (*e.g.,* diagrams) with textual descriptions, and providing domain-specific insights. The generalization is limited to specific domains, resulting in low-impact applications like data retrieval and integration.

C.2 Phase 2: Analogical Reasoning and Generalization

This stage emphasizes medium diversity and specificity in data, leveraging contextual understanding to enable relational and analogical reasoning. For instance, MLLMs may draw analogies between chemical reaction pathways and electrical circuits, enhancing interdisciplinary insights. The models exhibit medium-level generalization across domains, allowing for moderate complexity tasks like explaining scientific phenomena using cross-domain analogies.

C.3 Phase 3: Insightful Inference

As reasoning tasks grow in complexity, MLLMs in this phase focus on low diversity but high specificity datasets. Predictive reasoning mechanisms are central, enabling context-specific inferences. For example, a model might predict the behavior of a physical system under certain constraints or optimize complex processes like material design. These capabilities lead to high-impact applications, including scientific optimization and infer-

ential problem-solving.

C.4 Phase 4: Creative Hypothesis Generation

The final phase demands both high diversity and high creativity in data. Generative reasoning mechanisms empower MLLMs to propose innovative solutions and simulate hypotheses. For instance, models might design novel molecules for drug discovery or hypothesize ecological models for sustainable ecosystems. This phase represents very high-impact applications, bridging the gap between scientific discovery and innovation.

C.5 Summary

In summary, the Four Phases of Research Roadmap reflect the increasing complexity and potential of MLLMs in scientific reasoning. While current MLLMs have demonstrated impressive capabilities, they remain far from achieving AGI (Mumuni and Mumuni, 2025; Wang, 2024; Feng et al., 2024; Fei et al., 2022). The community must continue to explore advanced reasoning abilities, fostering collaboration and innovation along this roadmap to address the challenges of AGI-driven scientific reasoning.

D Data Differences among Four Domains

The reasoning capabilities of MLLMs make them particularly well-suited for processing heterogeneous multimodal data (Pattnayak et al., 2024; Li et al., 2024e,i). By integrating and analyzing diverse data formats such as text, images, and structured information, MLLMs enable more comprehensive insights across various scientific disciplines. This section highlights the unique data characteristics of four domains: mathematics, physics, chemistry, and biology, as summarized in Table 4.

D.1 Mathematics: Structured Abstraction

Mathematical data is characterized by its symbolic equations, graphs, and geometric figures. These data types are highly structured and abstract, requiring precise interpretation and manipulation. Visual features such as coordinate axes and geometric shapes often complement formal textual elements like equations and proofs. The integration of these modalities enables MLLMs to solve complex mathematical problems and support theorem proving (Chern et al., 2023; Yan et al., 2024a; Azerbayev et al., 2023; Li et al., 2024l).

D.2 Physics: Real-world Dynamics

Physics data encompasses diagrams (*e.g.,* vector and circuit diagrams), graphs, and descriptions of real-world phenomena. Its data structure reflects system dynamics and real-world applications, combining descriptive text with visual representations like force vectors or particle motion. MLLMs leverage these multimodal inputs to model physical systems and predict outcomes under varying conditions (Jaiswal et al., 2024; Barman et al., 2025; Yu et al., 2024).

D.3 Chemistry: Molecular and Symbolic

Chemistry relies heavily on molecular structures, reaction pathways, and the periodic table. The data is both symbolic and structural, involving visual elements such as 3D molecular models and reaction schemes. Textural features include chemical equations and reaction mechanisms. MLLMs facilitate understanding chemical interactions and even predicting new compounds or reaction outcomes (Alampara et al.; Miret and Krishnan, 2024; Mirza et al., 2024; Ramos et al., 2025; M. Bran et al., 2024).

D.4 Biology: Conceptual Complexity

Biological data is diverse, covering anatomical diagrams, cellular structures, and ecological models. Its structure is conceptual and often involves biological and anatomical representations. Visual inputs like cell diagrams and ecosystem models are paired with textual descriptions of processes and interactions. MLLMs support tasks such as identifying biological patterns and simulating ecological dynamics (Kraus et al., 2024; Luu and Buehler, 2023; Huang et al., 2024a; Wang et al., 2023a; Liu et al., 2023c; Tang et al., 2023).

D.5 Summary and Future Directions

In summary, MLLMs demonstrate strong reasoning capabilities across mathematics, physics, chemistry, and biology by integrating diverse multimodal data. However, their potential is not confined to these four disciplines. Future research should explore reasoning capabilities in broader domains such as geospatial analysis (Roberts et al., 2024; Mai et al., 2024; Hao et al., 2024; Zhong et al., 2024a) and coding (Li et al., 2024g; Guo et al., 2024a; Di et al., 2024; Hui et al., 2024), which demand advanced generalization abilities. Such endeavors are crucial for achieving the com-

Table 3: Comparison across four stages along key dimensions.

Dimension	Broad Knowledge and Recognition	Analogical Reasoning and Generalization	Insightful Inference	Creative Hypothesis Generation
Data and Knowledge Requirements	High diversity, low specificity	Medium diversity, medium specificity	Low diversity, high specificity	High diversity, high creativity
Reasoning Mechanisms	Low complexity (retrieval, alignment)	Medium complexity (relational, analogical)	High complexity (predictive)	Very high complexity (generative)
Model Generalization	Low (domain-specific)	Medium (cross-domain)	High (context-specific)	Very high (innovative)
Applications and Impact	Low impact (retrieval, integration)	Medium impact (interdisciplinary insights)	High impact (optimization, inference)	Very high impact (discovery, innovation)

prehensive reasoning capabilities envisioned in the AGI roadmap.

E Multi-Domain Scientific Reasoning Benchmark

The emergence of multi-domain scientific reasoning benchmarks has played a pivotal role in advancing AI models' ability to reason across diverse scientific domains. These benchmarks vary in their focus on multimodal integration, educational levels, and the comprehensiveness of domain coverage, as summarized in Table 5.

Comprehensive Domain Coverage: Several benchmarks, such as **MMMU-Pro**, **CMMMU**, and **SciEval**, provide extensive domain coverage, including mathematics, physics, chemistry, and biology. These benchmarks target diverse educational contexts, ranging from primary education to PhD-level tasks, ensuring broad applicability. Notably, **MMMU-Pro** and **SciBench** have become instrumental for college-level evaluation, while **CMMU** extends its scope to younger learners, addressing a critical gap in foundational scientific reasoning.

Multimodal Reasoning: The integration of multimodal capabilities has become a defining feature of contemporary benchmarks. Over 60% of the surveyed benchmarks, such as **EXAMS-V**, **SciBench**, and **ScienceQA**, incorporate multimodal tasks that involve textual, visual, and symbolic reasoning. These tasks reflect real-world problem-solving scenarios where multiple modalities interact, making them essential for evaluating the holistic reasoning capabilities of advanced AI models.

Specialized vs. General Benchmarks: While benchmarks like **OlympiadBench** and **OlympicArena** are tailored to high-stakes competitions, their narrow focus on specific tasks, such as Olympiad-level challenges, limits their generalizability. Conversely, resources like **AGIEval** and **SciEval** aim to provide a broader evaluation of scientific reasoning, covering multiple domains across various educational levels. However, their lack of multimodal integration highlights the need for more versatile benchmarks.

Future Directions: The future of multi-domain scientific reasoning benchmarks lies in developing unified frameworks that combine multimodal reasoning, domain comprehensiveness, and adaptability across educational levels. Incorporating elements like interactive feedback systems and collaborative reasoning tasks will bridge the gap between theoretical evaluation and practical applications, fostering the next generation of scientific reasoning models.

F Details of Hallucinations in Scientific Reasoning

Since a position paper must balance scope breadth and technical depth, we did not extensively elaborate on each specific challenge in the main text. However, we provide a more detailed discussion on hallucinations here, as they present a nuanced challenge particularly critical to the advancement of MLLMs in scientific reasoning. While often detrimental, understanding and managing hallucinations is key, especially considering the aspirational "Creative Hypothesis Generation" stage of our roadmap.

F.1 Different Types of Hallucinations in Scientific Reasoning

Hallucinations in MLLMs, broadly defined as the generation of information not grounded in the input data or established knowledge, manifest in several ways pertinent to scientific reasoning:

- **Factual Hallucination:** This is the most straightforward type, where the MLLM gener-

Table 4: Comparison of data features among four domains within the scope of this position paper.

Feature/Domain	Mathematics	Physics	Chemistry	Biology
Data Types	Symbolic equations, graphs, geometric figures	Diagrams (vector, circuit), graphs, real-world phenomena	Molecular structures, reaction pathways, periodic table	Anatomical diagrams, cellular structures, ecological models
Data Structure	Abstract, highly structured, formal notation	Real-world applications, system dynamics	Symbolic and molecular, structural	Conceptual, biological, and anatomical
Visual Features	Coordinate axes, geometric shapes	Diagrams e.g., force vectors, particle motion	3D molecular models, reaction schemes	Photos, cell diagrams, ecosystem models
Textural Features	Equations, proofs, formal definitions	Descriptive, experimental setups, theories	Chemical equations, reaction mechanisms	Biological processes, ecological interactions

Table 5: Comparison of multi-domain scientific reasoning benchmarks.

Paper	Organization	Venue	Multimodal	Education Level	Domain(s)			
					Math	Physics	Chemistry	Biology
SciVerse (Guo et al., 2025b)	CUHK	ACL Findings'25	✓	General		✓	✓	✓
MMSciBench (Ye et al., 2025)	Yale University	ACL Findings'25	✓	High	✓	✓		
EMMA (Hao et al., 2025)	UESTC	ICML'25	✓	General	✓	✓	✓	
CURIE (Cui et al., 2025)	Google	ICLR'25	✓	General		✓	✓	✓
MMVU (Zhao et al., 2025)	Yale University	CVPR'25	✓	College		✓	✓	✓
LLM-SRBench (Shojaee et al., 2025)	Virginia Tech	arXiv'25		General		✓	✓	✓
SciBench (Wang et al., 2024g)	UCLA	ICML'24	✓	College	✓	✓	✓	
MMMU (Yue et al., 2024a)	University of Waterloo & Ohio State University	CVPR'24	✓	College	✓	✓	✓	✓
EXAMS-V (Das et al., 2024)	MBZUAI	ACL'24	✓	General	✓	✓	✓	✓
ArxivQA/Cap (Li et al., 2024h)	HKU	ACL'24	✓	General	✓	✓	✓	✓
OlympiadBench (He et al., 2024a)	Tsinghua University	ACL'24	✓	Competition			✓	✓
SceMQA (Liang et al., 2024a)	University of Notre Dame	ACL Short'24	✓	College	✓	✓	✓	✓
SciEval (Sun et al., 2024b)	Shanghai Jiaotong University	AAAI'24		General	✓	✓	✓	✓
CMMU (He et al., 2024c)	Beijing Academy of Artificial Intelligence	IJCAI'24	✓	Primary/ Middle/High	✓	✓	✓	✓
AGIEval (Zhong et al., 2024b)	Microsoft	NAACL'24		High/College	✓	✓	✓	✓
CMMLU (Li et al., 2024d)	MBZUAI	ACL Findings'24		High/College	✓	✓	✓	✓
MMWorld (He et al., 2024b)	UCSC	arXiv'24	✓	General	✓	✓	✓	✓
MMMU-Pro (Yue et al., 2024b)	CMU	arXiv'24	✓	College	✓	✓	✓	✓
CMMMU (Zhang et al., 2024c)	HKUST	arXiv'24	✓	College	✓	✓	✓	✓
M4U (Wang et al., 2024a)	Chinese Academy of Sciences	arXiv'24	✓	College	✓	✓	✓	✓
MMSci (Li et al., 2024m)	UCSB	arXiv'24	✓	PhD		✓	✓	✓
OlympicArena (Huang et al., 2024b)	Shanghai Jiaotong University	arXiv'24		Competition	✓	✓	✓	✓
LitQA2 (Skarlinski et al., 2024)	FutureHouse Inc.	arXiv'24		General			✓	✓
ScienceAgentBench (Chen et al., 2024c)	OSU	arXiv'24		General			✓	✓
JEEBench (Arora et al., 2023)	Microsoft & UC Berkeley	EMNLP'23		General	✓	✓	✓	
M3Exam (Zhang et al., 2023)	Alibaba	NeurIPS'23	✓	Primary/ Middle/High	✓		✓	
LitQA (Lála et al., 2023)	FutureHouse Inc.	arXiv'23		General			✓	✓
ScienceQA (Lu et al., 2022a)	UCLA	NeurIPS'22	✓	Primary/High		✓	✓	✓
IconQA (Lu et al., 2021b)	UCLA	NeurIPS'21	✓	General	✓	✓		
MMLU (Hendrycks et al., 2021)	UC Berkeley	ICLR'21		General	✓	✓	✓	✓

ates incorrect or non-existent facts, such as an incorrect physical constant, a misremembered chemical property, a wrong date for a discovery, or a fabricated citation. In science, where precision is paramount, factual hallucinations can lead to fundamentally flawed conclusions.

- **Conceptual Hallucination:** This involves the misrepresentation or misapplication of scientific concepts, theories, or principles. An MLLM might, for example, confuse correlation with causation in a biological context, misapply Newton's laws in a scenario where relativistic effects are dominant, or incorrectly define a mathematical theorem's conditions.

- **Relational Hallucination:** This occurs when an MLLM fabricates or misrepresents relationships between entities, variables, or concepts. It might invent a non-existent interaction between two proteins, incorrectly link a gene to a disease without evidence, or propose a flawed causal chain in a physical process.

- **Explanatory Hallucination:** Here, the MLLM generates explanations for phenomena that are either logically unsound, inconsistent with established theories, or based on fabricated intermediate steps, even if the final conclusion appears plausible. This is particularly dangerous as it can create a false sense of understanding. For instance, an MLLM might "derive" a correct math-

ematical formula through a series of incorrect algebraic manipulations.

- **Exploratory Hallucination:** This is a more nuanced category. It refers to the generation of novel, plausible-sounding hypotheses or ideas that are not yet validated or directly supported by existing data. While technically a "hallucination" (as it's not grounded in current certainties), this type can be beneficial in the "Creative Hypothesis Generation" stage if properly managed and clearly identified as speculative. The risk lies in presenting such exploratory outputs as established facts.

F.2 Existing Mitigation Strategies

Several strategies have been developed to mitigate hallucinations in general MLLM applications, which can be adapted, with limitations, to scientific reasoning:

- **Knowledge Retrieval Augmentation (RAG):** This involves grounding the MLLM's responses by retrieving relevant information from external, trusted knowledge bases (*e.g.,* scientific databases, curated literature). The model then uses this retrieved context to formulate its answer, reducing reliance on its parametric memory which can be a source of hallucinations.

- **Consistency Checks:**
  - *Self-consistency:* Generating multiple reasoning paths or answers for the same query and selecting the most consistent or frequently occurring one.
  - *Cross-modal consistency:* For MLLMs, ensuring that information derived from text aligns with information from images (*e.g.,* a diagram in a physics problem should match the textual description).
  - *Logical consistency:* Evaluating the logical coherence of the generated reasoning steps.

- **Process-Based Supervision (and Process Reward Models - PRMs):** Instead of only rewarding the correctness of the final answer, these methods involve supervising or rewarding the intermediate steps of the reasoning process. This encourages the model to follow valid inferential pathways, making it less likely to "jump" to conclusions through hallucinated steps.

- **Human Feedback (RLHF/RLAIF):** Reinforcement Learning from Human Feedback (and its variants like RLAIF) involves training the MLLM based on human ratings of its outputs. Humans can penalize hallucinated content, guiding the model towards more factual and reliable responses.

F.3 Why Current Mitigation Methods Fall Short in Scientific Reasoning

Despite their utility, current hallucination mitigation strategies face significant hurdles when applied to the rigorous demands of scientific reasoning:

- **Complexity and Nuance of Scientific Knowledge:**
  - Scientific knowledge is vast, deeply interconnected, and often highly nuanced or domain-specific. RAG systems might struggle to retrieve the *exact* precise piece of information needed or may misinterpret the context of retrieved documents. The sheer volume can also lead to retrieving conflicting information.
  - Many scientific concepts are abstract and symbolic (*e.g.,* in quantum mechanics or advanced mathematics), making it difficult for models to truly "understand" and ground them, even with retrieved text.

- **The Need for Exploration and Controlled Creativity:**
  - Aggressive hallucination mitigation, designed to enforce strict factuality, can inadvertently stifle the model's ability to engage in exploratory reasoning or generate novel hypotheses (critical for Stage 4). The challenge is to allow for "beneficial" exploratory generation while suppressing detrimental factual or conceptual hallucinationsa balance current methods struggle to achieve.

- **Lack of Granular Explainability and Verifiability:**
  - Even if a mitigation technique reduces overt hallucinations, the internal reasoning process of MLLMs often remains a black box. For scientific applications, it's not enough for an answer to be correct; the reasoning must be transparent, verifiable, and

align with established scientific methods. Current methods don't inherently guarantee this level of scrutiny.
  - Error propagation in multi-step scientific reasoning is a major issue. A subtle, uncaught hallucination in an early step can invalidate the entire chain, and PRMs might not be sufficiently fine-grained to catch all such nuanced errors in complex scientific domains.

- **Data Limitations for Scientific Domains:**
  - High-quality, large-scale, and *verified* multimodal scientific datasets specifically curated for training MLLMs to avoid domain-specific hallucinations are scarce.
  - Training effective PRMs or RLAIF systems for science requires extensive, expert-annotated data on correct reasoning processes, which is costly and time-consuming to produce across diverse scientific fields.
  - Models trained on existing scientific literature might inherit biases or outdated information present in that corpus, leading to "hallucinations" relative to the current state of knowledge.

In conclusion, while existing mitigation techniques provide a starting point, the unique demands of scientific reasoningits need for precision, depth, verifiability, and controlled creativitynecessitate the development of more sophisticated, domain-aware, and interpretable approaches to manage and understand hallucinations in MLLMs.

G Multimodal Scientific (M)LLMs

Scientific (M)LLMs represent an emerging class of models that integrate multiple modalities to tackle complex problems across various scientific disciplines, as summarized in Table 6. These models leverage massive datasets combining text, images, graphs, and other forms of scientific data to provide deeper insights and advanced reasoning capabilities. Unlike traditional models, which typically focus on either textual or visual information, scientific MLLMs are designed to harmonize and synthesize diverse sources of scientific knowledge, enabling enhanced performance in tasks such as mathematical problem solving, chemical reaction prediction, biological analysis, and physical simulations (Morris, 2023; Pei et al., 2024; Reddy and Shojaee, 2024; Zhang et al., 2024f).

H Discussion of More Modalities in Scientific Reasoning

While current MLLMs have demonstrated remarkable progress, their sensory input is predominantly confined to text and static images. This limitation curtails their ability to fully comprehend and reason about the multifaceted nature of scientific phenomena, which often unfold dynamically, possess intricate three-dimensional structures, or generate non-visual sensory data. To significantly advance scientific reasoning, MLLMs must evolve to integrate a richer spectrum of modalities, including audio, video, 3D structural data, and other sensor-derived information. This expansion is not merely about processing more data types but about enabling a more holistic and nuanced understanding essential for complex scientific inquiry.

H.1 Current Limitations and the Imperative for Richer Modalities

The current text-image paradigm in MLLMs, while powerful, presents several inherent limitations when applied to the diverse needs of scientific reasoning:

- **Restricted Modality Spectrum:** Most MLLMs excel with visual and textual data but struggle to incorporate or meaningfully reason over audio (*e.g.,* sonification of data, acoustic signatures of experiments), video (*e.g.,* dynamic processes, experimental procedures), and 3D data (*e.g.,* molecular configurations, geological strata). This restricts their ability to capture the full context of many scientific observations.

- **Inadequate for Dynamic Phenomena:** Scientific processes are often dynamic. For instance, observing a chemical reaction, the growth of a biological culture, or the time-evolution of a physical system requires processing temporal information embedded in video or sequential sensor readings. Static images and text alone cannot capture these crucial dynamic aspects.

- **Insufficient for Spatial Complexity:** Many scientific domains, such as chemistry (molecular structures), biology (protein folding, anatomical systems), materials science (crystal lattices), and earth sciences (geophysical models), fundamentally rely on understanding complex 3D spatial relationships. Current MLLMs often lack the sophisticated 3D geometric reasoning capabilities required.

- **Limited Cross-Modal Association for Complex Reasoning:** Scientific reasoning frequently involves drawing inferences across multiple, diverse data streams. For example, correlating a change in a visual indicator (image/video) with a specific sound (audio) and a corresponding data log (text/tabular) to deduce a causal link in an experiment demands finer-grained and more diverse modality interactions than currently supported.

- **Domain-Specific Knowledge Integration Challenges:** Effectively interpreting specialized modalities (*e.g.,* spectrographic data, sonified outputs) necessitates not just raw data processing but also deep integration with domain-specific ontologies and knowledge bases to contextualize the sensory input accurately.

H.2 Technical Approaches for Broader Modality Integration

Addressing these limitations requires developing robust technical approaches for encoding and integrating a wider array of modalities, tailored to their unique characteristics and scientific relevance:

- **Audio Processing:**
  - *Feature Extraction:* Leveraging established techniques like spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) to convert raw audio signals from experiments (*e.g.,* equipment sounds, material stress responses) or sonified data into feature representations suitable for neural networks.
  - *Model Adaptation:* Developing specialized audio encoders or adapting existing ones, and aligning their latent spaces with those of text and image encoders through cross-modal attention mechanisms or joint embedding strategies. This allows, for instance, an MLLM to associate the sound of a failing component with its visual depiction and textual description of failure modes.

- **Video Understanding:**
  - *Spatiotemporal Modeling:* Employing 3D Convolutional Neural Networks (3D CNNs), Video Transformers, or other video-language models to capture both spatial features within frames and temporal dynamics across frames. This is critical for analyzing experimental procedures, observing cellular motility, or tracking particle trajectories.
  - *Keyframe and Event Detection:* Implementing mechanisms to automatically identify and focus on critical events or keyframes within a scientific video (*e.g.,* phase transition, initiation of a reaction), thereby reducing computational load and highlighting salient information for reasoning.

- **3D Structural Data Processing:**
  - *Geometric Encoding:* Utilizing encoders like PointNet, Point Transformer, or Graph Neural Networks (GNNs) to process irregular 3D data such as point clouds (*e.g.,* from LiDAR scans of geological sites) or molecular meshes.
  - *Incorporating Physical and Chemical Constraints:* Enhancing structural understanding by integrating domain-specific knowledge, such as bond lengths and angles in molecular modeling or material properties in engineering design, directly into the encoding or reasoning process. This ensures that interpretations are scientifically plausible.

- **Sensor and Time-Series Data:**
  - *Specialized Encoders:* Using Recurrent Neural Networks (RNNs), 1D CNNs, or Transformers adapted for time-series data from various sensors (*e.g.,* temperature, pressure, EEG, spectroscopy).
  - *Temporal Alignment:* Developing methods to align asynchronous and heterogeneous sensor streams with other modalities like video or textual logs of experimental events.

H.3 Advanced Cross-Modal Interaction and Scientific Reasoning

The true power of incorporating more modalities lies in enabling more sophisticated cross-modal interactions and reasoning capabilities:

- **Unified Multimodal Alignment:**

- *Shared Embedding Space:* Designing a common embedding space, potentially through CLIP-style contrastive learning extended to audio, video, and 3D data, where features from diverse modalities representing the same scientific concept are brought closer together.
- *Dynamic Modality Weighting:* Implementing mechanisms that allow the model to dynamically assign importance (weights) to different modalities based on the specific scientific task, context, or data quality, ensuring robustness and flexibility.

- **Enhanced Scientific Reasoning Modules:**
  - *Multimodal Causal Graph Generation:* Constructing more comprehensive causal relationship graphs by leveraging evidence from diverse inputs. For example, observing a video of an experimental setup, listening to audio cues indicating a process stage, and reading instrument outputs (text/tabular) can collectively inform the nodes and edges of a causal graph explaining an observed phenomenon (*e.g.,* "increased gas pressure [sensor] caused by heating element activation [video/log] led to audible valve release [audio]").
  - *Generative Reasoning with Richer Evidence:* Empowering language models to generate more nuanced and well-supported scientific explanations, hypotheses, or experimental designs by integrating and citing evidence from audio logs, video segments, 3D visualizations, and sensor data alongside textual knowledge.

H.4 Task-Driven Training Strategies and Future Directions

Realizing the potential of MLLMs with expanded modalities requires innovative training strategies and a forward-looking research agenda:

- **Multi-Task Learning for Comprehensive Understanding:**
  - Training models on a combination of tasks, including fundamental modality understanding (*e.g.,* audio event classification, video action recognition, 3D object recognition) alongside higher-level scientific reasoning tasks (*e.g.,* predicting experimental outcomes, inferring material properties from multimodal sensor data).

- **Leveraging Domain-Specific Knowledge Graphs:**
  - Incorporating expert-annotated scientific knowledge graphs (ontologies, relational databases) as prior constraints or auxiliary information during training. This can guide the model in interpreting complex multimodal inputs within the correct scientific framework (*e.g.,* linking a 3D protein structure to its known functions and interaction pathways).

- **Addressing Data Scarcity:**
  - Developing techniques for data augmentation, synthetic data generation (*e.g.,* simulating experimental videos or 3D molecular interactions), and leveraging weakly supervised or self-supervised learning methods, given that large-scale, richly annotated multimodal scientific datasets are often scarce.

- **Future Outlook:** The path forward involves creating benchmark datasets that encompass these richer modalities within scientific contexts. Furthermore, research should explore how these expanded sensory capabilities can enable truly interactive MLLMs that can actively participate in the scientific process, for instance, by suggesting modifications to an experimental setup based on real-time video and sensor feedback. Such advancements will be pivotal in unlocking new frontiers in AI-assisted scientific discovery.

I Discussion of Alternative Views

While this paper champions the significant potential of MLLMs to advance scientific reasoning, it is crucial to engage with pertinent alternative viewpoints and concerns that challenge this optimism. Acknowledging these perspectives allows for a more nuanced understanding of the MLLM adoption pathway in science.

I.1 Domain-Specific Models as a Superior Alternative

One argument suggests that *highly specialized, domain-specific models tailored to individual scientific disciplines may outperform general-purpose MLLMs in reasoning tasks*, as indicated in Figure 6 (a). Proponents of this view highlight

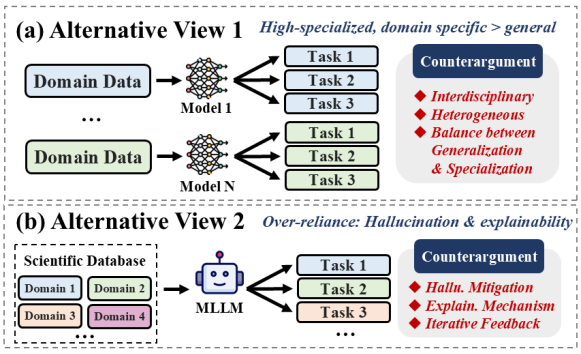

Figure 6: Illustrations of alternative view 1 (a) and view 2 (b), as well as our corresponding counterarguments.

that scientific reasoning often requires deep domain expertise, nuanced understanding, and customized data processing pipelines that are difficult to replicate in generalized multimodal architectures (Zhang et al., 2024d; Barman et al., 2025). For example, domain-specific models like AlphaFold for protein structure prediction (Jumper et al., 2021) or symbolic computation tools for solving mathematical problems (Mirzadeh et al., 2024; Lu et al., 2021a) excel precisely because of their narrow focus.

Counterargument. While domain-specific models have demonstrated remarkable success in narrow applications, they lack the flexibility to generalize across multiple domains or integrate diverse modalities of data. Scientific reasoning increasingly involves interdisciplinary approaches, such as bioinformatics combining biology and computational techniques, or climate science requiring both geospatial and textual analysis (Reddy and Shojaee, 2024). MLLMs, by design, offer the ability to process and reason over heterogeneous data sources, enabling a broader and more integrative approach (Zhang et al., 2024f). Furthermore, domain-specific knowledge can still be fine-tuned within MLLMs, allowing these systems to leverage the best of both generalization and specialization (Chen et al., 2023c).

More Elaboration can be seen as follows:

- **Necessity of Customized Architectures and Data Handling:**
  - *What creates the necessity of a customized model?* Many scientific domains rely on unique, highly structured data formats (*e.g.,* SMILES strings in chemistry, FASTA sequences in genomics, PDB files for protein structures, specific notations in physics diagrams or mathematical proofs). General MLLMs, even with multimodal capabilities, may lack the inherent inductive biases or fine-grained understanding to parse, interpret, and reason over these specialized representations with the required precision. A custom model can be built from the ground up to "speak" the native language of that data.
  - Domain-specific models can embed established physical laws, chemical reaction rules, or biological pathways more explicitly, either through architectural design or by training on curated datasets that heavily emphasize these principles. This contrasts with MLLMs that must infer these rules implicitly from broader, potentially noisier data.
  - The sheer precision required for tasks like predicting quantum mechanical properties, drug-target interaction affinities, or complex engineering tolerances might necessitate models whose entire objective function and training data are geared towards minimizing error in that specific, narrow domain, rather than general plausibility.
- **Knowledge Depth and Nuance:**
  - Scientific breakthroughs often require profound depth of knowledge and nuanced understanding within a very specific sub-field. While MLLMs aim for breadth, they might only achieve a superficial understanding in areas requiring years of human expert training. The "curse of dimensionality" applies to knowledge as well; covering all scientific domains deeply in one model is an immense challenge.
  - Training datasets for general MLLMs, even if including scientific papers, might be diluted by vast amounts of non-scientific text, or may not capture the long tail of niche, but critical, scientific knowledge that specialized databases and models incorporate.

While the counterargument in the main paper highlights MLLM flexibility for interdisciplinary tasks, proponents of specialization argue that many critical scientific problems are bottlenecked by depth in a single discipline, not breadth. They contend that fine-tuning an MLLM might offer some domain adaptation, but it's unlikely to match a model purpose-built for the intricacies of that domain.

I.2 Risks of Over-reliance on MLLMs

Another valid concern is *the potential over-reliance on MLLMs, which could exacerbate issues such as hallucination and lack of explainability*, as shown in Figure 6 (b). Critics argue that MLLMs, while powerful, are prone to generating plausible-sounding but incorrect or unsubstantiated outputs. This risk is particularly concerning in scientific reasoning, where accuracy and rigor are paramount (Bai et al., 2024b). Additionally, the black-box nature of these models makes it difficult for researchers to validate their reasoning processes or trust their conclusions, which could hinder their adoption in critical scientific applications (Cambria et al., 2024; Rodis et al., 2024; Dang et al., 2024b).

Counterargument. These concerns are valid and underscore the need for a cautious and measured approach to MLLM adoption. However, rather than dismissing MLLMs outright, these issues highlight areas for improvement. For example, integrating explainability mechanisms, such as visual attention maps (Chefer et al., 2021; Dehimi and Tolba, 2024) or rationale generation (Hu and Yu, 2024; Wu et al., 2024a), can enhance transparency. Additionally, hybrid models that combine MLLMs with symbolic reasoning (Li et al., 2024a; Zhou et al., 2024a) or expert systems (Guan et al., 2024; Niu et al., 2024) can mitigate risks while maintaining strengths of multimodal reasoning. Finally, iterative feedback loops and human-in-the-loop systems can ensure reliability in reasoning workflows (Xiao et al., 2024; Zou et al., 2024; Zheng et al., 2024b).

More Elaboration can be seen as follows:

- **The Amplified Peril of Scientific Hallucination:**
  - Hallucinations in a general chatbot might be an annoyance; in science, they can be dangerous or resource-wasting. This isn't just about generating factually incorrect statements, but also subtle misinterpretations of complex multimodal inputs (*e.g.,* misreading a crucial detail in a micrograph or circuit diagram) that lead to flawed reasoning chains.
  - The "plausible-sounding but incorrect" nature of MLLM outputs is especially insidious in science, where a subtly flawed equation, a non-existent chemical reaction pathway, or an incorrectly inferred biological interaction could derail research or lead to unsafe experimental protocols.
- **The Imperative of Explainability and Verifiability:**
  - The scientific method fundamentally relies on transparency, reproducibility, and the ability to scrutinize the reasoning process. The "black-box" nature of many MLLMs directly conflicts with this. How does one verify the multi-step reasoning an MLLM used to arrive at a novel hypothesis from a complex set of inputs (*e.g.,* a research paper, experimental data, and a diagram)?
  - Current explainability techniques (*e.g.,* attention maps, rationale generation) often provide superficial or post-hoc justifications that may not reflect the true internal "reasoning" of the model, if such a coherent process even exists in a human-understandable form. This makes debugging errors or building trust in novel, MLLM-generated insights exceptionally difficult.
- **Bias Propagation and Deskilling Concerns:**
  - MLLMs trained on existing scientific literature and datasets will inevitably inherit and potentially amplify existing biases within those sources (*e.g.,* overrepresentation of certain research topics, demographic groups in clinical data, or established theories at the expense of emerging ones).
  - A more philosophical concern is the potential for "deskilling." If researchers become overly reliant on MLLMs for hypothesis generation, data interpretation, or even experimental design, there's a risk of atrophying fundamental human scientific reasoning and critical thinking skills, reducing scientists to sophisticated prompt engineers.

The main paper's counterargument focuses on mitigation strategies. However, critics would argue that the current state of these mitigations is insufficient for high-stakes scientific applications, and the burden of proof lies in demonstrating consistent reliability, deep interpretability, and bias control before widespread adoption in critical scientific endeavors. The very nature of

MLLMs—learning statistical patterns rather than causal mechanisms—might pose a fundamental limit to their trustworthiness in a domain that prizes causal understanding.

J Clarification of LLM Usage

In the preparation of this manuscript, the authors leveraged Gemini-Pro-2.5 as an assistive writing tool. The use of the LLM was strictly confined to an editorial capacity, primarily for refining language, improving grammatical correctness, and enhancing the clarity and readability of the text. The core intellectual contributionsincluding the central thesis, the proposed four-stage research roadmap, the analysis of challenges, and the overall structure of the paperwere conceived and developed exclusively by the human authors. The authors meticulously reviewed, edited, and approved all final wording to ensure it accurately reflects their original ideas and arguments, and bear full responsibility for the accuracy and integrity of the content.

Table 6: Summary of scientific (M)LLMs.

Paper	Organization	Venue	Multimodal	Parameter	Domain(s)			
					Math	Physics	Chemistry	Biology
General-Reasoner (Ma et al., 2025)	Waterloo			4B/7B/14B	✓	✓	✓	✓
OmniScience (Prabhakar et al., 2025)	SES AI			70B	✓	✓	✓	✓
Galactica (Taylor et al., 2022)	Meta		✓	125M/1.3B/6.7B/ 30B/120B	✓	✓	✓	✓
LLM-SR (Shojaee et al., 2024)	Virginia Tech			8x7B	✓	✓		✓
SciLitLLM (Li et al., 2024k)	USTC			7B/14B		✓	✓	✓
Darwin series (Xie et al., 2023)	University of New South Wales			7B		✓	✓	
SPMM (Chang and Ye, 2022)	KAIST	Nature Communications'24	✓	–			✓	✓
InstructMol (Cao et al., 2023b)	IDEA & HKUST	COLING'25	✓	7B			✓	✓
BioinspiredLLM (Luu and Buehler, 2023)	MIT	Advanced Science'24		13B			✓	✓
nach0 (Livne et al., 2023)	NVIDIA	Chemical Science'24	✓	250M/780M			✓	✓
Mole-BERT (Xia et al., 2023)	Westlake University	ICLR23		–			✓	✓
BioReason (Fallahpour et al., 2025)	University of Toronto			1B/4B				✓
BioGPT (Luo et al., 2022)	Microsoft	Briefings in Bioinformatics'22		355M				✓
Evolla (Zhou et al., 2025)	Westlake University		✓	80B				✓
Prollama (Lv et al., 2024)	Peking University			7B				✓
Perovskite-LLM (Liu et al., 2025c)	HKUST(GZ)	EMNLP Findings'25		8B			✓	
ether0 (Narayanan et al., 2025)	FutureHouse Inc			24B			✓	
Chemdfm (Zhao et al., 2024c)	Shanghai Jiaotong University			13B			✓	
ChemDFM-X (Zhao et al., 2024b)	Shanghai Jiaotong University		✓	8B			✓	
ChemLLM (Zhang et al., 2024a)	Shanghai AI Lab			7B			✓	
GIT-Mol (Liu et al., 2023c)	Peng Cheng Lab	Computers in Biology and Medicine'24	✓	700M			✓	
MolGPT (Bagal et al., 2021)	International Institute of Information Technology	Journal of Chemical Information and Modeling'21		6M			✓	
MOOSE-Chem (Yang et al., 2024f)	NTU & Shanghai AI Lab			–			✓	
BatGPT-Chem (Yang et al., 2024d)	Shanghai Jiaotong University			15B			✓	
DARWIN 1.5 (Xie et al., 2024)	University of New South Wales			7B			✓	
MolMetaLM (Wu et al., 2024b)	Central South University			–			✓	
SMI-TED (Soares et al., 2024)	IBM			289M			✓	
MathCoder (Wang et al., 2023c)	CUHK	ICLR'24		7B/13B	✓			
MAmmoTH1 (Yue et al., 2023)	UWaterloo	ICLR24		7B/13B/70B	✓			
Math-LLaVA (Shi et al., 2024b)	NUS	EMNLP Findings'24		13B	✓			
JiuZhang 2.0 (Zhao et al., 2023b)	RUC & iFLYTEK	KDD'23		–	✓			
JiuZhang 1.0 (Zhao et al., 2022)	RUC & iFLYTEK	KDD'22		145M	✓			
Minerva (Lewkowycz et al., 2022)	Google	NeurIPS22		8B/62B/540B	✓			
Hypertree Proof Search (Lample et al., 2022)	Meta	NeurIPS'22		–	✓			
Qwen2.5-Math (Yang et al., 2024b)	Alibaba			1.5B/7B/72B	✓			
Qwen2-Math (Yang et al., 2024a)	Alibaba			1.5B/7B/72B	✓			
Qwen2-Math-Instruct (Yang et al., 2024a)	Alibaba			1.5B/7B/72B	✓			
MathGPT (TALEducation, 2023)	TAL Group		✓	130B	✓			
math-specialized Gemini 1.5 Pro (Team et al., 2024b)	Google		✓	–	✓			
InternLM2-Math (Ying et al., 2024)	Shanghai AI Lab			1.8B/7B/20B/ 8x22B	✓			
InternLM2.5-StepProver (Wu et al., 2024c)	Shanghai AI Lab			7B	✓			
Llemma (Azerbayev et al., 2023)	Princeton University & Eleuther AI			7B/34B	✓			
ChatGLM-Math (Xu et al., 2024b)	Zhipu AI			32B	✓			
MetaMath (Yu et al., 2023)	Cambridge & Huawei			7B/13B/70B	✓			
MathGLM (Yang et al., 2023b)	Tsinghua & Zhipu AI			10M/100M/335M/ 500M/2B/6B/10B	✓			
MathGLM-Vision (Yang et al., 2024e)	Tsinghua & Zhipu AI		✓	9B/19B/32B	✓			
Skywork-13B-Math (Yang et al., 2023a)	SkyworkAI		✓	7B/13B	✓			
DeepSeekMath (Shao et al., 2024)	DeepSeek AI			7B	✓			
DeepSeekProver-V1 (Xin et al., 2024a)	DeepSeek AI			7B	✓			
DeepSeek-Prover-V.15 (Xin et al., 2024b)	DeepSeek AI			7B	✓			
Mathstral (Mistral AI Team, 2024)	Mistral AI			7B	✓			
JiuZhang 3.0 (Zhou et al., 2024b)	RUC & iFLYTEK			7B/8B	✓			
Math-LLM (Liu et al., 2024c)	East China Normal University		✓	8.26B/7B/72B	✓			
GPT-f (Polu and Sutskever, 2020)	OpenAI			160M/400M/700M/	✓			
Rho-Math (Lin et al., 2024)	Microsoft			1B/7B	✓			
WizardMath (Luo et al., 2023)	Microsoft			7B/70B	✓			
Xwin-LM (Ni et al., 2024)	Microsoft			7B/13B/70B	✓			
MAmmoTH2 (Yue et al., 2024c)	UWaterloo			7B/8B	✓			
GAIRMath-Abel (Chern et al., 2023)	Shanghai Jiaotong University			7B/13B/70B	✓			
KwaiYiiMath (Fu et al., 2023)	Kuaishou			13B	✓			
k0-math (MoonshotAI, 2024)	Moonshot AI			–	✓			
NuminaMath (Beeching et al., 2024)	Numina			7B/72B	✓			