

# SCIVERSE: Unveiling the Knowledge Comprehension and Visual Reasoning of LMMs on Multi-modal Scientific Problems

Ziyu Guo\* Renrui Zhang\* Hao Chen\* Jialin Gao\*  
Dongzhi Jiang Jiaze Wang Pheng-Ann Heng†

The Chinese University of Hong Kong  
ziyuguo@link.cuhk.edu.hk

## Abstract

The rapid advancement of Large Multi-modal Models (LMMs) has enabled their application in scientific problem-solving, yet their fine-grained capabilities remain under-explored. In this paper, we introduce SCIVERSE, a multi-modal scientific evaluation benchmark to thoroughly assess LMMs across 5,735 test instances in five distinct versions. We aim to investigate three key dimensions of LMMs: *scientific knowledge comprehension*, *multi-modal content interpretation*, and *Chain-of-Thought (CoT) reasoning*. To unveil whether LMMs possess sufficient scientific expertise, we first transform each problem into three versions containing different levels of knowledge required for solving, i.e., Knowledge-free, -lite, and -rich. Then, to explore how LMMs interpret multi-modal scientific content, we annotate another two versions, i.e., Vision-rich and -only, marking more question information from texts to diagrams. Comparing the results of different versions, SCIVERSE systematically examines the professional knowledge stock and visual perception skills of LMMs in scientific domains. In addition, to rigorously assess CoT reasoning, we propose a new scientific CoT evaluation strategy, conducting a step-wise assessment on knowledge and logical errors in model outputs. Our extensive evaluation of different LMMs on SCIVERSE reveals critical limitations in their scientific proficiency and provides new insights into future developments. Project page: <https://sciverse-cuhk.github.io>.

## 1 Introduction

In recent years, the rapid advancement of large models, i.e., Large Language Models (LLMs) (OpenAI, 2023a; Touvron et al., 2023a,b; Chiang et al., 2023) and Large Multi-modal Models (LMMs) (Liu et al., 2023b; OpenAI, 2023c; Zhang

et al., 2024b; Gao et al., 2024; Zong et al., 2024), has significantly expanded the frontiers of various modalities and scenarios, such as text (OpenAI, 2023b, 2024b; Guo et al., 2025a), 2D images (OpenAI, 2023c; Zhang et al., 2023; Zhu et al., 2023; Zhang et al., 2024a), and 3D point clouds (Guo et al., 2024; Xu et al., 2023; Guo et al., 2023; Jia et al., 2024). Notably, LMMs have demonstrated promising potential in addressing multi-modal scientific problems across diverse domains, including physics, chemistry, and biology.

Despite efforts to develop scientific datasets with visual content as evaluation benchmarks (Lu et al., 2022; Yue et al., 2023, 2024), existing approaches primarily assess LMMs through basic testing, where models directly solve original problems and are compared based on overall accuracy. However, we identify that effective problem-solving in this domain requires three key skills: *scientific knowledge comprehension*, *multi-modal content interpretation*, and *Chain-of-Thought (CoT) reasoning*. Consequently, the fine-grained scientific capabilities of LMMs remain insufficiently explored, lacking a detailed and thorough examination within the research community.

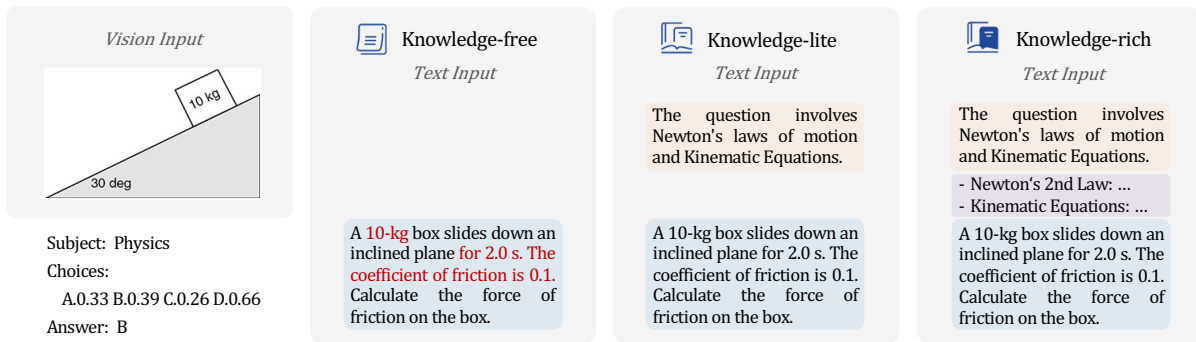
In this paper, we introduce SCIVERSE, a comprehensive evaluation benchmark to assess LMMs on multi-modal scientific problems. Our curated dataset comprises 1,147 meticulously collected problems and 5,735 newly annotated test instances across five distinct versions, covering difficulty levels from high school to college. Specifically, to investigate the three key skills aforementioned, we aim to explore the following questions regarding scientific problem-solving as outlined in Figure 1.

1. *Do LMMs possess sufficient scientific knowledge to solve the problems?* Unlike general visual scenarios, scientific problem-solving requires LMMs to have prior knowledge of specific subjects. Previous benchmarks do

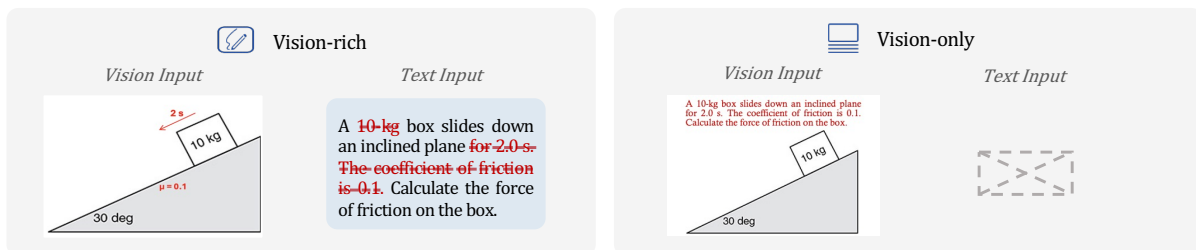
\* Equal Contributions † Corresponding Author

## Scientific Knowledge Comprehension

Given Condition + Core Question    Knowledge Cue    Knowledge Detail



## Multi-modal Content Interpretation



## Scientific CoT Evaluation Strategy

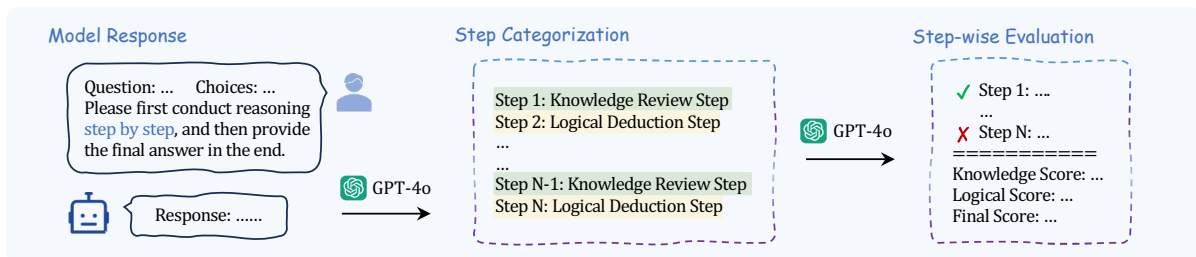


Figure 1: **Overview of Five Problem Versions and our Scientific CoT Evaluation Strategy in SCIVERSE.** To unveil the scientific knowledge comprehension (Top), we first transform each problem into three versions integrating different levels of expertise knowledge. Then, to examine the multi-modal content interpretation (Middle), we further annotate two problem versions with varying vision-language information. We introduce a specialized scientific evaluation strategy (Bottom) to assess the fine-grained reasoning capabilities of LMMs.

not differentiate between errors caused by a lack of knowledge and deficiencies in logical reasoning. To address this, we manually transform each problem of SCIVERSE into three versions with increasing levels of embedded knowledge within question texts: Knowledge-free, Knowledge-lite, and Knowledge-rich. By exposing LMMs to different depths of domain expertise, we systematically analyze how knowledge comprehension impacts scientific problem-solving.

2. **Can LMMs effectively interpret question information from multi-modal content?** In existing benchmarks, problem conditions are primarily presented in textual form, enabling

LMMs to process them through language modeling. However, in real-world scenarios, key information is often embedded in diagrams, or even the entire question is printed as visual input (e.g., scanned documents, handwritten notes, or screenshots). Thus, it is essential to evaluate how LMMs perform when question content is progressively shifted from text to visual modalities. To this end, we further annotate the problems in SCIVERSE into two additional versions: Vision-rich and Vision-only. These versions systematically measure LMMs' perception and OCR capabilities to retrieve and process multi-modal contexts in scientific problems.

### 3. *Is CoT reasoning effective in improving the accuracy of solving scientific problems?*

Rather than directly providing a final answer, Chain-of-Thought (CoT) reasoning breaks the problem-solving process into a sequence of logical steps. In the context of scientific problems, the intermediate steps typically fall into two categories: knowledge review and logical deduction. Existing benchmarks generally assess CoT performance based on direct answer accuracy or a binary ‘True’ or ‘False’ metric. In contrast, we propose a new scientific CoT evaluation strategy using GPT-4o (OpenAI, 2024a). Our approach first extracts key steps from the model’s output and then performs a step-wise analysis, identifying both knowledge and reasoning errors. This methodology offers a more comprehensive evaluation of the CoT reasoning capabilities of LMMs.

With five curated problem versions and a detailed CoT evaluation, our benchmark challenges LMMs to demonstrate not only expert knowledge but also their ability to integrate and reason across multiple modalities under varying levels of complexity. We evaluate a wide range of popular LMMs on SCIVERSE, offering unique insights to the research community. Our findings reveal that closed-source LMMs outperform open-source LMMs in both knowledge comprehension and visual perception in scientific domains. However, both categories of models struggle with Vision-only problems, which resemble real-world scenarios. Additionally, closed-source models exhibit stronger CoT reasoning capabilities, producing higher-quality reasoning steps.

Our contributions are threefold:

- We present **SCIVERSE**, a multi-modal evaluation benchmark specifically designed to assess scientific reasoning across various disciplines. For the first time, SCIVERSE highlights three critical challenges that LMMs face in scientific problem-solving.
- We develop a set of five problem versions that target distinct scientific reasoning challenges, addressing previous evaluation limitations in knowledge comprehension and multi-modal interpretation of LMMs.
- We introduce a scientific CoT evaluation strategy, focusing on step-wise errors in both

knowledge review and reasoning deduction. This approach offers a comprehensive analysis of LMMs’ scientific CoT capabilities.

## 2 SCIVERSE

In Section 2.1, we first present an overview of SCIVERSE, including dataset statistics and the collection process. Then, we respectively illustrate our methodology on the three critical aspects of assessing LMMs: scientific knowledge comprehension (Section 2.2), multi-modal content interpretation (Section 2.3), and Chain-of-Thought (CoT) reasoning evaluation (Section 2.4).

### 2.1 Dataset Overview

To comprehensively evaluate scientific reasoning, we curate a diverse set of problems spanning a wide range of disciplines and knowledge domains.

**Data Statistics.** Table 1 and Figure 2 provide an overview of the key statistics and subject distribution of SCIVERSE. The dataset consists of 5,735 problems, divided across three major domains: Physics, Chemistry, and Biology. These subjects are further broken down into 21 distinct scientific topics, allowing for an evaluation of problem-solving performance at a granular level. SCIVERSE includes five different problem versions, each consisting of 1,147 instances, designed to assess both the knowledge expertise and visual perception capabilities of LMMs. With more knowledge content integrated, the question length, from Knowledge-free, Knowledge-lite, to Knowledge-rich versions, also increases. As the information is gradually transitioned from texts to diagrams, the question length decreases from Knowledge-lite, Vision-rich, to Vision-only versions.

**Data Curation.** To guarantee a comprehensive scope, we begin by reviewing publicly available scientific datasets, from which we curate an initial set of 1,200 problems sourced from three datasets: SceMQA (Liang et al., 2024), MMMU (Yue et al., 2023), and CMMM (Ge et al., 2024). To maintain high quality, we engage eight PhD-level science experts to carefully evaluate and select problems based on the knowledge complexity and visual richness of problems. Subsequently, we translate all texts into English in a Latex format, and convert the problem types into multiple-choice questions. After a thorough review process, 1,147 problems are retained, each of which is then transformed into

Statistic	Number
Total questions	5,735
Questions of each version	1,147
<i>Knowledge-free</i>	
Maximum question length	1,353
Average question length	254.3
<i>Knowledge-lite</i>	
Maximum question length	1,991
Average question length	491.6
<i>Knowledge-rich</i>	
Maximum question length	2,768
Average question length	842.2
<i>Vision-rich</i>	
Maximum question length	1,239
Average question length	227.5
<i>Vision-only</i>	
Maximum question length	0
Average question length	0

Table 1: **Key Statistics of SCIVERSE.**

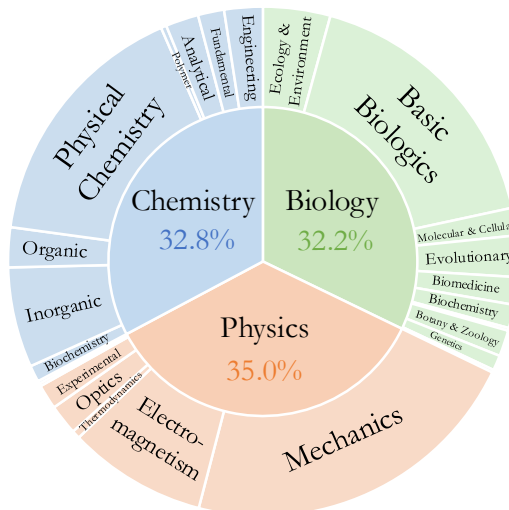


Figure 2: **Subject Distribution of SCIVERSE.** The dataset contains 2,010 questions from Physics, 1,880 from Chemistry, and 1,845 from Biology.

five different versions, as outlined in the following sections and illustrated in Figure 3.

## 2.2 Scientific Knowledge Comprehension

A key challenge for LMMs in solving scientific problems is their capability to comprehend sufficient domain knowledge, which is essential for understanding the question and performing multimodal reasoning. To evaluate this, we manually transform each problem in SCIVERSE into three versions, each incorporating varying levels of scientific knowledge. By comparing the performance of an LMM across these three versions, we aim to investigate the impact of knowledge comprehension on scientific problem-solving.

**Knowledge-free Version.** We first eliminate all background knowledge from the question text, leaving only the core question, which includes the given condition (e.g., *slides down for 2.0 s*) and core question (e.g., *calculate the force*). This version presents a significant challenge for LMMs, as they must first interpret the question accurately and then relate it to the appropriate scientific knowledge for problem-solving. The content in both text and visual modalities is structured as follows:

Text Input: Given Condition + Core Question

Vision Input: Diagram

**Knowledge-lite Version.** Based on the previous version, we introduce a simple knowledge cue in

the question text, indicating the high-level knowledge required for solving the problem. Typically, we provide related theorem names or formulation references at the beginning of the question, such as *Newton’s laws of motion* or *Kinematic Equations*. These cues help guide the LMMs in interpreting the problem and allow us to assess whether their performance improves when provided with basic background knowledge, compared to Knowledge-free results. The content is structured as:

Text Input: Knowledge Cue + Given Condition + Core Question

Vision Input: Diagram

**Knowledge-rich Version.** In this version, we further enrich the problem with detailed scientific information, such as specific equations and the application method of a relevant theorem (e.g., *“This law states that the net force ( $F$ ) acting on an object is equal to the product of its mass ( $m$ ) and its acceleration ( $a$ ).”*). By comparing performance in the Knowledge-rich and -lite versions, we can determine whether LMMs truly comprehend the expertise required and whether their performance improves when provided with more detailed background information. The content is structured as:

Text Input: Knowledge Cue + Knowledge Detail + Given Condition + Core Question

Vision Input: Diagram



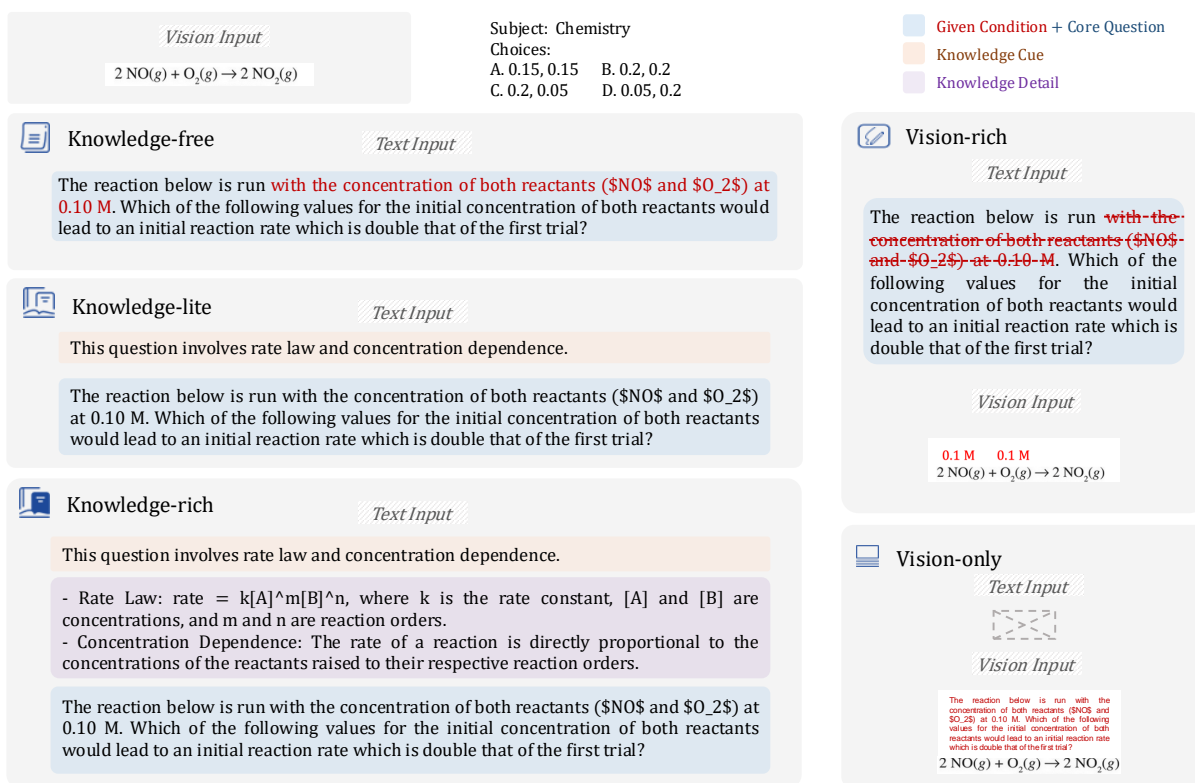


Figure 3: **Examples of Five Problem Versions in SCIVERSE.** For each problem in SCIVERSE, we first create the Knowledge-free version by removing all knowledge content from the question text. Next, we add knowledge cues and details to produce the Knowledge-lite and Knowledge-rich versions. Additionally, starting from the Knowledge-free version, we generate two more versions, Vision-rich and Vision-only, where the given condition and, ultimately, the entire question are transferred to the visual diagram.

### 2.3 Multi-modal Content Interpretation

In contrast to LLMs, LMMs must accurately interpret the diagram input and integrate visual information with the textual question for effective question-answering. Therefore, we focus on evaluating the visual perception capabilities of LMMs in the context of scientific diagrams. To this end, we transform each problem in SCIVERSE into two additional versions that progressively shift the balance of question information from text to diagrams, which are more similar to real-world scenarios.

**Vision-rich Version.** On top of the Knowledge-lite version, we remove most of the problem conditions from the question text (e.g., “slides down for 2.0 s”) and instead annotate them directly onto the diagram, provided they can be suitably represented visually. This version challenges LMMs to rely more on the visual modality for extracting critical information, reducing the reliance on textual content and testing their true multi-modal interpretation capabilities in scientific problem-solving. The content is structured as follows:

Text Input: Core Question

Vision Input: Diagram + Given Condition

**Vision-only Version.** In this version, we take the integration of visual information a step further by embedding the entire question directly onto the diagram, eliminating any textual input. This setup closely mirrors real-world scenarios where users capture an image or screenshot of a problem. Without any textual cues, vision-only problems present the most challenging evaluation for LMMs, which assess their capabilities in knowledge comprehension, OCR, and visual perception. The content is structured as follows:

Text Input: None

Vision Input: Diagram + Given Condition + Core Question

### 2.4 Scientific CoT Evaluation Strategy

For complex scenarios, utilizing CoT (Wei et al., 2022) to perform step-by-step reasoning is essential for improving the problem-solving accuracy of LMMs. While some previous scientific benchmarks (Yue et al., 2024) have reported the CoT performance, they still rely on a binary ‘True’ or

*Text Input*

**Question:**  
The acceleration versus time of a bicycle rider is shown here: Assuming that the bicycle starts from the origin at an initial velocity of +4 m/s, complete the following questions: Determine the total distance traveled during the 5 s of motion.

**Choice:**  
A. 38.5    B. 29    C. 42.5    D. 31.5

*Vision Input*

### Step-wise Evaluation

Knowledge Review Step
  Logical Deduction Step

- ✓ To solve this problem, we need to apply kinematic principles.
- ✓ First, from t = 0 to t = 5 seconds, the acceleration is not zero, so the velocity keeps changing.
- ✓ To find the change in velocity over each interval, we use the formula  $\Delta v = a \times t$ .
- ✓ The velocity at t = 2 is:  $v_1 = v_0 + \Delta v = v_0 + at$ .
- ✗ For t = 0 to t = 2, the acceleration is +1 m/s<sup>2</sup>, so  $v_1 = 4 + 1 \times 2 = 8$  m/s.
- ✗ Now, let's calculate the distance using the equation  $d_1 = v_1 \times t$ .
- ✗ Therefore,  $d_1 = 6 \times 2 = 12$  m.
- ✓ Now calculate the second-interval distance using the displacement equation  $d = v_0t + (1/2)at^2$ .
- ✗ For t = 2 to t = 5, the initial velocity is 12 m/s, acceleration is -3 m/s<sup>2</sup>, and the time is 3 s:  $d_2 = 12 \times 3 + (1/2) \times (-3) \times 3^2 = 30 - 4.5 = 22.5$  m.
- ✗ Finally, the total distance traveled is:  $d_{\text{total}} = d_1 + d_2 = 12 + 22.5 = 31.5$  m.
- ✓ Thus, the total distance traveled is 31.5 meters, which corresponds to D. 31.5.

Figure 4: **Examples of the Scientific CoT Evaluation Strategy.** For reasoning responses from LMMs, we prompt GPT-4o (OpenAI, 2024a) to perform two evaluation stages, i.e., step categorization and step-wise evaluation. We categorize the intermediate steps into two types: knowledge review and logical reasoning.

‘False’ metric based on the final answer, overlooking the quality of the intermediate steps during reasoning. To address this gap, we propose a specialized scientific CoT evaluation strategy designed to assess the fine-grained CoT capabilities of LMMs in scientific problem-solving. This strategy involves two sequential stages using GPT-4o (OpenAI, 2024a) as shown in Figure 4.

**Step Categorization.** For model responses generated using CoT prompting (Kojima et al., 2022), we first apply GPT-4o to extract the key steps from the extended reasoning sequence and categorize them into two types:

- **Knowledge Review Step** refers to the process

of quoting or recalling relevant expert knowledge during problem-solving (e.g., “we need to apply kinematic principles”). These review steps assist LMMs in subsequent reasoning but may be prone to errors, such as quoting an irrelevant theorem or misremembering equations.

- **Logical Deduction Step** involves applying logical reasoning to derive an intermediate or final conclusion, which can be either a calculated result (e.g., “ $d_1 = 6 \times 2 = 12$  m”) or a knowledge-based inference (e.g., “so the velocity keeps changing”). This step may encounter errors, such as incorrect calculations, improper substitutions, or flawed inferences.

Model	All		Knowledge-rich				Knowledge-free				Knowledge-lite				Vision-rich				Vision-only			
	Acc	Sci-CoT	Acc	Sci-CoT	Sci-CoT <sub>K</sub>	Sci-CoT <sub>L</sub>	Acc	Sci-CoT	Sci-CoT <sub>K</sub>	Sci-CoT <sub>L</sub>	Acc	Sci-CoT	Sci-CoT <sub>K</sub>	Sci-CoT <sub>L</sub>	Acc	Sci-CoT	Sci-CoT <sub>K</sub>	Sci-CoT <sub>L</sub>	Acc	Sci-CoT	Sci-CoT <sub>K</sub>	Sci-CoT <sub>L</sub>
Baseline																						
Random Chance	22.7	-	22.7	-	-	-	22.7	-	-	-	22.7	-	-	-	22.7	-	-	-	22.7	-	-	-
Closed-source LLMs																						
GPT-4V	45.7	52.3	47.1	55.8	72.3	39.3	46.8	54.1	69.3	41.8	46.6	52.9	66.4	39.4	46.0	52.0	65.2	38.8	42.1	50.7	60.4	41.0
Gemini-1.5-Pro	49.5	58.6	50.8	62.2	78.2	46.2	50.7	60.9	76.3	45.5	50.5	58.4	70.7	46.1	49.9	57.3	68.4	46.2	45.9	55.2	64.3	46.1
Claude-3.5-Sonnet	52.8	62.4	54.1	66.9	80.2	53.6	53.9	63.4	78.8	48.0	53.7	62.5	75.3	49.7	53.1	61.3	72.5	50.1	49.3	59.3	69.9	48.7
GPT-4o	54.0	66.7	55.3	70.8	84.6	57.0	55.2	67.8	80.3	55.3	55.0	66.4	78.2	54.6	54.4	66.4	76.3	56.5	50.2	64.0	71.4	56.6
Open-source LLMs																						
SPHINX-Tiny (1.1B)	27.6	30.2	28.9	34.7	38.2	31.2	29.1	31.4	34.4	28.4	26.7	29.8	31.4	28.2	26.1	29.5	30.9	28.1	27.2	26.5	28.2	24.8
MiniGPT-v2 (7B)	30.0	34.1	30.5	37.8	41.2	34.4	29.6	35.7	38.2	33.2	31.5	39.9	34.0	45.8	31.6	32.8	33.1	32.5	26.9	30.3	30.9	20.7
ShareGPT4V (13B)	33.4	36.9	36.3	41.3	44.6	38.0	34.7	37.2	41.3	33.1	32.3	36.9	38.7	35.1	32.3	36.7	37.1	36.3	31.5	32.5	33.9	31.1
LLaVA-1.5 (13B)	33.7	38.2	35.4	42.8	45.1	40.5	35.0	39.5	41.9	37.1	32.7	38.2	39.7	36.7	33.0	37.9	39.2	36.6	32.3	34.0	34.9	33.1
LLaVA-NeXT (8B)	36.4	39.4	39.0	43.1	48.2	38.0	39.1	40.6	46.8	34.4	36.6	39.0	43.1	34.9	36.1	39.3	42.3	36.3	31.3	35.6	38.2	33.0
InternLM-XC2 (7B)	36.7	40.9	39.9	44.4	49.3	39.5	38.8	41.8	46.3	37.3	37.6	40.6	44.2	37.0	35.5	40.6	43.0	38.2	31.6	37.8	40.2	55.6
SPHINX-MoE (8x7B)	37.3	41.1	41.3	44.0	48.7	39.3	38.8	41.9	47.2	36.6	38.9	39.9	43.7	36.1	36.3	41.3	44.2	38.4	31.4	38.2	41.1	35.3
SPHINX-Plus (13B)	37.3	41.2	41.4	44.5	49.1	39.9	38.7	42.0	47.6	36.4	39.2	41.0	44.9	37.1	37.4	41.1	44.8	37.4	29.6	38.4	41.0	35.8
InternVL-1.5 (26B)	39.0	46.3	40.4	49.8	51.3	48.3	41.7	47.3	49.9	44.7	39.5	46.2	49.2	43.2	39.2	46.1	48.8	43.4	34.3	42.5	41.3	43.7
InternVL-2 (8B)	42.6	49.9	43.9	53.8	59.2	48.4	43.7	50.7	58.3	43.1	43.2	49.7	54.3	45.1	42.9	49.7	52.9	46.5	39.1	46.1	49.2	43.0
Qwen2-VL (7B)	44.7	53.2	46.2	57.3	63.1	51.5	45.9	54.0	63.2	44.8	45.8	53.3	61.3	45.3	45.0	53.1	61.7	44.5	40.4	48.6	60.0	37.2
LLaVA-OneVision (7B)	46.1	51.3	47.6	54.6	61.7	47.5	47.2	54.0	61.4	46.6	47.0	51.1	60.3	41.9	46.5	49.7	59.9	39.5	41.9	47.3	59.7	34.9

Table 2: **Evaluation Results on Five Problem Versions of SCIVERSE.** The ‘All’ scores represent the average results across all five problem versions. The metric ‘Acc’ refers to the binary ‘True’ or ‘False’ evaluation based solely on the final answer. ‘Sci-CoT’ refers to our proposed scientific CoT evaluation strategy, averaging the scores of knowledge review and logical reasoning, denoted as ‘Sci-CoT<sub>K</sub>’ and ‘Sci-CoT<sub>L</sub>’. The highest scores for closed-source and open-source LLMs are marked in red and blue, respectively.

**Step-wise Evaluation.** Following the step categorization, we prompt GPT-4o to provide a fine-grained ‘True’ or ‘False’ judgment for each individual step. This step-wise evaluation thoroughly considers each intermediate step, offering insights into the detailed CoT reasoning capabilities of LLMs. Subsequently, we compute two average scores: one for the knowledge comprehension steps and another for the logical deduction steps. In contrast to the previous binary accuracy, our strategy, which generates two distinct scores, provides a more comprehensive assessment of the model’s understanding of scientific knowledge and its proficiency in CoT reasoning.

### 3 Experiment

In Section 3.1, we first introduce our experimental settings, including the evaluation LLMs and implementation details. Then, in Section 3.2, we provide the performance comparison and insightful analysis on SCIVERSE.

#### 3.1 Evaluation Settings

**Evaluation Models.** We comprehensively assess a wide range of open-source and closed-source LLMs on SCIVERSE. Closed-source models include Gemini-1.5-Pro (Gemini Team, 2023), Claude-3.5-Sonnet (Anthropic, 2024), GPT-4V (OpenAI, 2023c), and GPT-4o (OpenAI, 2024a). Open-source models include MiniGPT-v2 (Chen et al., 2023a), LLaVA-1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024), LLaVA-

OneVision (Li et al., 2024b), ShareGPT4V (Chen et al., 2023b), SPHINX series (Gao et al., 2024), InternLM-XComposer-2 (Dong et al., 2024), InternVL-1.5 (Chen et al., 2024a), InternVL-2 (Chen et al., 2024a), Qwen2-VL (Qwen Team, 2024), and Qwen2.5-VL (Team, 2025).

**Implementation Details.** We adopt two metrics for evaluation. The first is the previous binary metric solely based on the final answer, termed ‘Acc’. We adopt an input prompt, “*directly provide the answer*”, to guide LLMs to provide the final answer directly. The second is our proposed scientific CoT evaluation strategy. We term the scores of knowledge and logical errors as ‘Sci-CoT<sub>K</sub>’ and ‘Sci-CoT<sub>L</sub>’, respectively, and denote their average score as ‘Sci-CoT’. We adopt an input CoT prompt, “*perform reasoning step-by-step*”, to elicit step-wise reasoning output. We evaluate all LLMs in a zero-shot setting without few-shot examples. We also provide a baseline representing random chance by randomly selecting an option. All evaluation is conducted on NVIDIA A100 GPUs.

#### 3.2 Discussion and Analysis

In Table 2, we present the detailed evaluation results of SCIVERSE. Based on the performance comparison, we derive several key observations:

- As more knowledge is provided, open-source LLMs show greater improvement, whereas closed-source LLMs exhibit relatively smaller gains. As we move from the Knowledge-free,

Knowledge-lite, to Knowledge-rich versions, most LMMs demonstrate performance improvements as more knowledge cues and details are added to the question. Among these, closed-source LMMs, such as GPT-4o and Claude-3.5-Sonnet, display relatively stable results across all three versions. This stability suggests that these models inherently possess a greater depth of expertise knowledge and are better able to effectively leverage it for problem-solving. This trend is further supported by the results of ‘Sci-CoT<sub>K</sub>’, where closed-source models achieve higher accuracy in knowledge review compared to their open-source counterparts.

- *When more information is shifted to vision input, open-source LMMs experience a significantly larger performance drop compared to closed-source LMMs.* From Knowledge-free to Vision-rich versions, most LMMs exhibit a noticeable performance decline. This suggests that, relative to text-based question information, LMMs face greater challenges when problem conditions are given as visual information. Such results highlight the limitations in the visual encoding quality and cross-modal understanding of current LMMs when applied to scientific diagrams. Additionally, closed-source LMMs show a smaller performance drop between the two problem versions, indicating their relatively stronger capabilities in scientific visual perception.
- *The most challenging scenario for LMMs occurs with Vision-only problems, where all question information is embedded in diagrams.* The largest performance drop is observed between the Vision-rich and Vision-only versions for both closed-source and open-source LMMs. This indicates that LMMs struggle with low capabilities for the OCR and interpretation of question information embedded visually in diagrams. Such a lack of reliable OCR capabilities and cross-modal integration severely hinders LMMs’ potential to tackle scientific problems in real-world scenarios.
- *Closed-source LMMs demonstrate notably stronger CoT reasoning capabilities than open-source LMMs.* When comparing the ‘Acc’ and ‘Sci-CoT’ scores across all prob-

lem versions, we observe a significant gap, with the CoT evaluation score being higher than the binary accuracy. This suggests that many intermediate steps may be correct, even when the final answer is incorrect. Such cases would be overlooked by the traditional binary accuracy metric, but our scientific CoT evaluation strategy effectively identifies and incorporates them into the final scores. Furthermore, the gap between the two scores is more pronounced in closed-source LMMs, indicating that closed-source models excel at CoT reasoning, producing higher-quality intermediate steps and more robust overall performance.

## 4 Related Work

### 4.1 Multi-modal Scientific Benchmark

Recent advances in LMMs have sparked significant interest in their mathematic (Zhang et al., 2024c,d) and scientific reasoning capabilities, particularly in tasks involving visual interpretation. A spectrum of scientific benchmarks has emerged across different educational levels: ScienceQA (Lu et al., 2022) targets elementary and secondary education, focusing on foundational scientific concepts. Moving to higher education, SceMQA (Liang et al., 2024) introduces a comprehensive benchmark at the college entrance level, encompassing Mathematics, Physics, Chemistry, and Biology. At the collegiate level, MMMU (Yue et al., 2023) and its enhanced version MMMU-Pro (Yue et al., 2024) have emerged as broader benchmarks, spanning diverse fields from arts to technology. The multilingual expansion is demonstrated by CMMMMU (Ge et al., 2024), which extends the evaluation framework to Chinese contexts. For advanced evaluation, OlympiadBench (He et al., 2024) incorporates challenging Mathematics and Physics Olympiad problems, testing LMMs’ capabilities in solving exceptionally difficult problems. Meanwhile, some recent works (Guo et al., 2025b; Jiang et al., 2025, 2024) also focus on the exploration of the *Retrieval-Augmented Generation (RAG)* ability and *Chain-of-Thought (CoT) reasoning* reasoning ability of the LMMs. Different from all previous works, our SCIVERSE, for the first time, investigate three critical issues within LMMs in scientific problem-solving, i.e., *scientific knowledge comprehension*, *multi-modal content interpretation*, and *Chain-of-Thought (CoT) reasoning*, offering unique insights to the community.



## 4.2 Large Multi-modal Models (LMMs)

Recent advances in multi-modal AI have been marked by significant developments in LMMs, which combine the capabilities of LLMs and vision models to process diverse visual inputs. While proprietary models like GPT-4V (OpenAI, 2023c), Claude (Anthropic, 2024), Gemini (Gemini Team, 2023), and GPT-4o (OpenAI, 2024a) have shown remarkable visual reasoning abilities, their closed nature has spurred the development of open-source alternatives. Early open-source LMMs like LLaVA (Liu et al., 2023b) and MiniGPT-4 (Zhu et al., 2023) paired CLIP-based image encoders (Radford et al., 2021) with LLMs for multi-modal instruction tuning. Later models such as LLaVA-NeXT (Li et al., 2024a), LLaVA-OneVision (Li et al., 2024b), ShareGPT4V (Chen et al., 2023b), InternVL (Chen et al., 2024b), SPHINX (Lin et al., 2023), and Qwen-VL (Qwen Team, 2024) expanded these capabilities through broader training datasets and advanced training strategies. In this paper, we aim to comprehensively evaluate their fine-grained capabilities in scientific domains, guiding the future developments of LMMs.

## 5 Conclusion

In this paper, we introduce SCIVERSE, a comprehensive multi-modal benchmark designed to evaluate the fine-grained capabilities of LMMs in scientific problem-solving. By transforming problems into multiple versions that vary in knowledge and modality, we investigate three critical dimensions of LMMs: scientific knowledge comprehension, multi-modal content interpretation, and CoT reasoning. Furthermore, our proposed scientific CoT evaluation strategy provides a deeper understanding of how LMMs handle knowledge and logical errors during problem-solving. The findings from our extensive evaluation of current state-of-the-art LMMs underscore the need for further advancements in their scientific proficiency and multi-modal reasoning capabilities. Moving forward, we hope SCIVERSE may serve as a foundation for future developments of LMMs in scientific fields.

## Acknowledgements

The work described in this paper was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CUHK 14200824; and by the Hong Kong

## Limitations

Although our primary focus is on investigating the three critical issues of LMMs in scientific domains, rather than the breadth of evaluation, future work could expand SCIVERSE to include additional disciplines and scenarios, such as art, business, medicine, and social sciences.

## References

- Anthropic. 2024. Claude-3.5. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Jun Chen, Deyao Zhu<sup>1</sup> Xiaoqian Shen<sup>1</sup> Xiang Li, Zechun Liu<sup>2</sup> Pengchuan Zhang, Raghuraman Krishnamoorthi<sup>2</sup> Vikas Chandra<sup>2</sup> Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.
- Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv*, abs/2311.12793.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.
- Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.
- Zhang Ge, Du Xinrun, Chen Bei, Liang Yiming, Luo Tongxu, Zheng Tianyu, Zhu Kang, Cheng Yuyang, Xu Chunpu, Guo Shuyue, Zhang Haoran, Qu Xingwei, Wang Junjie, Yuan Ruibin, Li Yizhi, Wang Zekun, Liu Yudong, Tsai Yu-Hsuan, Zhang Fengji, Lin Chenghua, Huang Wenhao, Chen Wenhui, and Fu Jie. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.20847*.
- Google Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025b. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*.
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. 2024. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.
- Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, et al. 2024. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. *arXiv preprint arXiv:2411.18623*.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. 2025. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, et al. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. <https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/>.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.
- 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. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- OpenAI. 2023a. Chatgpt. <https://chat.openai.com>.
- OpenAI. 2023b. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- OpenAI. 2023c. [GPT-4V\(ision\) system card](#).
- OpenAI. 2024a. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2024b. [Introducing openai o1, 2024](#).
- Qwen Team. 2024. Qwen2-vl. *Qwen Team*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Qwen Team. 2025. [Qwen2.5-vl](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2023. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2024. [Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark](#). *Preprint*, arXiv:2409.02813.
- Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. 2023. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*.
- Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024a. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR 2024*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and

Yu Qiao. 2024b. [LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention](#). In *The Twelfth International Conference on Learning Representations*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024c. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. 2024d. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. 2024. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*.



## A Detailed Annotation Process

### A.1 Creation of Five Problem Versions

Our detailed annotation process is as follows:

**1) Initial Construction.** All five versions are manually constructed by eight PhD-level science experts, with GPT-4o (OpenAI, 2024a) used as auxiliary tools for efficiency and consistency. The annotators design the content transformations, including knowledge cues, detailed scientific information, and vision-based annotations, based on pedagogical and domain-specific considerations.

**2) Refinement and Review.** After annotation, we employ GPT-4o and Claude 3.5 (Anthropic, 2024) to iteratively review all instances. This step focuses on cleaning noise, removing any unintended answer leakage from the question stem, and ensuring formatting consistency (valid LaTeX syntax). This human-in-the-loop pipeline ensures both content quality and structural rigor across all versions.

### A.2 Reliable Answer Correctness

For all the questions, we directly adopt the ground-truth answers provided in the original datasets, and unify all questions into a multi-choice, single-answer format to ensure consistency across the benchmark. Specifically, we handle different original formats as follows:

1. If the original format is *multi-choice single-answer*, we directly use the original question and answer. Annotators conduct basic checks for option clarity, LaTeX formatting, and potential noise.
2. If the original format is *multi-choice multi-answer*, we reformulate them into single-answer questions by refining the question wording and choices to ensure a unique correct answer, thus aligning with the benchmark’s overall structure.
3. If the original format is *free-form*, the annotators construct answer candidates based on domain knowledge, with optional assistance from GPT for generating distractors and verifying solutions.

All modified questions and generated choices were double-checked by PhD-level experts to ensure correctness, uniqueness, and consistency with the original intent and ground truth.

## B GPT-4o Prompt for the CoT Scientific Evaluation Strategy

### B.1 Step Categorization

You will be provided with a step-by-step solution to a problem. Your task is to:

1. Break the solution into the smallest possible steps, ensuring each step represents a single action or piece of reasoning.
2. Classify each step as either:
  - $\{K\}$ : **Knowledge Review Step** (facts, definitions, or prior knowledge used in the step).
  - $\{L\}$ : **Logical Deduction Step** (deductions, calculations, or inferences made in the step).

### B.2 Step-wise Evaluation

You will be provided with a list of steps from a solution for a scientific question, each classified as either Knowledge Review Step ( $\{K\}$ ) or Logical Deduction Step ( $\{L\}$ ). Your task is to assign a correctness score to each step:

- $\{1\}$ : **Correct** (the knowledge is relevant, sufficient, and accurate, or the reasoning is logically valid).
- $\{0\}$ : **Incorrect** (the knowledge is irrelevant, insufficient, or inaccurate, or the reasoning is flawed).

## C Human Study

To verify the reliability of our step-wise evaluation conducted using GPT-4o (OpenAI, 2024a), we conduct a human study comparing GPT’s annotations with expert judgments.

Specifically, we randomly select 30 questions (150 instances in total across all five versions) and extract the step-by-step outputs from LLaVA-OneVision (7B) (Li et al., 2024b). Five PhD-level science experts independently annotate each reasoning step (classified as either Knowledge Review Step or Logical Deduction Step), resulting in 926 annotated steps totally. We then compared these annotations with those generated by GPT-4o and found a **97.1%** agreement rate (**899 among 926**), indicating strong alignment between model-based and human evaluations.

## D Additional Examples

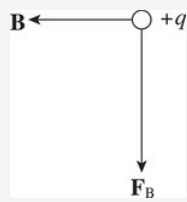
In Figures 5~13, we provide more examples of different problem versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Physics


Choices:

- A. Downward, in the plane of the page
- B. Upward, in the plane of the page
- C. Out of the plane of the page
- D. Into the plane of the page

 Knowledge-free

*Text Input*


In the figure below, what must be the direction of the particle's velocity,  $v$ ?

 Knowledge-lite

*Text Input*

The right-hand rule is a mnemonic for understanding direction of the magnetic force.

In the figure below, what must be the direction of the particle's velocity,  $v$ ?

 Knowledge-rich

*Text Input*

The right-hand rule is a mnemonic for understanding direction of the magnetic force.

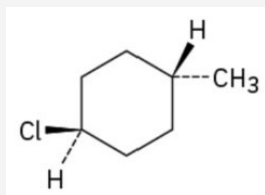
For a positive charge:  
Point your fingers in the direction of the velocity  $v$ .  
Curl your fingers towards the direction of the magnetic field  $\mathbf{B}$ .  
Your thumb will point in the direction of the magnetic force  $\mathbf{F}_B$ .

In the figure below, what must be the direction of the particle's velocity,  $v$ ?

Figure 5: Examples of Different Problem Versions in SCIVERSE.

Core Question Knowledge Cue Knowledge Detail

*Vision Input*



Subject: Chemistry

Choices:

- A. cis-1-Chloro-4-methylcyclohexane
- B. trans-1-Chloro-4-methylcyclohexane
- C. 1-Chloro-4-ethylcyclohexane
- D. 2-Chloro-4-methylcyclohexane

Knowledge-free

*Text Input*

Name the following substances, including the cis- or trans- prefix.

Knowledge-lite

*Text Input*

This question involves the naming of cycloalkanes.

Name the following substances, including the cis- or trans- prefix.

Knowledge-rich

*Text Input*

This question covers the naming of cycloalkanes.

It involves counting carbon atoms, locating substituents, and assigning prefixes. The cis-trans notation specifies the spatial arrangement of substituents, crucial for understanding compound properties.

Name the following substances, including the cis- or trans- prefix.

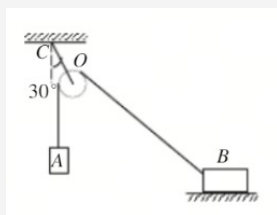
Figure 6: Examples of Different Problem Versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Physics


Choices:

- A. 23
- B. 26
- C. 45
- D. 34.6

 Knowledge-free

*Text Input*


As shown in the figure, rope CO is at an angle of 30 degrees to the vertical direction, O is a fixed pulley, and objects A and B are connected by a thin rope across the fixed pulley, in a state of equilibrium. It is known that the mass of B is 10 kg, and the support force of the ground on B is 80 N. If the size of the pulley is not considered, the friction between object B and the ground is \_\_\_N?

 Knowledge-lite

*Text Input*

This question requires Static Equilibrium, Rope tension and Components of force.

As shown in the figure, rope CO is at an angle of 30 degrees to the vertical direction, O is a fixed pulley, and objects A and B are connected by a thin rope across the fixed pulley, in a state of equilibrium. It is known that the mass of B is 10 kg, and the support force of the ground on B is 80 N. If the size of the pulley is not considered, the friction between object B and the ground is \_\_\_N?

 Knowledge-rich

*Text Input*

This question requires Static Equilibrium, Rope tension and Components of force.

- Static equilibrium: When an object is in equilibrium (not moving), the sum of all the forces acting on it must equal zero. This means that the upward and downward forces on the object must exactly cancel each other out.
- Rope tension: In a fixed pulley system, the tension in the rope is uniform throughout the rope. In an ideal situation with no friction and no mass in the pulley, the tension in the rope would be equal on both sides of the pulley.
- Components of force: The tension with a certain degree has components in the vertical direction and in the horizontal direction. These components can be calculated using trigonometric functions, especially sine and cosine.

As shown in the figure, rope CO is at an angle of 30 degrees to the vertical direction, O is a fixed pulley, and objects A and B are connected by a thin rope across the fixed pulley, in a state of equilibrium. It is known that the mass of B is 10 kg, and the support force of the ground on B is 80 N. If the size of the pulley is not considered, the friction between object B and the ground is \_\_\_N?

Figure 7: Examples of Different Problem Versions in SCIVERSE.

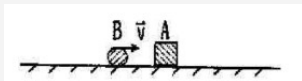


Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Physics


Choices:

- A.  $(mv)^2 / (m+m)^2 g\mu$
- B.  $(mv)^2 / 2(m+m)^2 g\mu$
- C.  $(mv)^2 / 3(m+m)^2 g\mu$
- D.  $(mv)^2 / 4(m+m)^2 g\mu$

 Knowledge-free

*Text Input*


As shown in the figure, a mass  $M$  object  $A$  is at rest on a horizontal surface with a coefficient of kinetic friction  $\mu$ . Another object  $B$  with mass  $m$  collides completely inelastically with object  $A$  while moving horizontally to the right with velocity  $v$ . The horizontal distance  $L$  they slide after the collision is \_\_\_\_.

 Knowledge-lite

*Text Input*

The scenario described involves a completely inelastic collision between two objects, Object A and another unnamed object, on a horizontal plane. In physics, a completely inelastic collision is one where the colliding objects stick together after impact, losing all relative kinetic energy in the process.

As shown in the figure, a mass  $M$  object  $A$  is at rest on a horizontal surface with a coefficient of kinetic friction  $\mu$ . Another object  $B$  with mass  $m$  collides completely inelastically with object  $A$  while moving horizontally to the right with velocity  $v$ . The horizontal distance  $L$  they slide after the collision is \_\_\_\_.

 Knowledge-rich

*Text Input*

The scenario described involves a completely inelastic collision between two objects, Object A and another unnamed object, on a horizontal plane. In physics, a completely inelastic collision is one where the colliding objects stick together after impact, losing all relative kinetic energy in the process.

Conservation of momentum is a fundamental principle in collisions, stating that the total momentum of a system remains constant, regardless of internal forces acting within the system. Additionally, the concept of kinetic energy, which is the energy associated with motion, is crucial in understanding the energy loss during inelastic collisions.

As shown in the figure, a mass  $M$  object  $A$  is at rest on a horizontal surface with a coefficient of kinetic friction  $\mu$ . Another object  $B$  with mass  $m$  collides completely inelastically with object  $A$  while moving horizontally to the right with velocity  $v$ . The horizontal distance  $L$  they slide after the collision is \_\_\_\_.

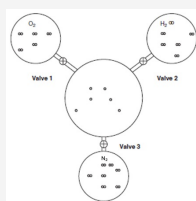
Figure 8: Examples of Different Problem Versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Chemistry


Choices:

- A. 1.0 atm 0.5 atm
- B. 0.41 atm 0.82 atm
- C. 0.81 atm 1.65 atm
- D. 2.0 atm 1.0 atm

 Knowledge-free

*Text Input*


Three 1-liter flasks are connected to a 3-liter flask by valves. The 3-liter flask is evacuated to start and the entire system is at 585 K. The first flask contains oxygen, the second hydrogen, and the third nitrogen. The pressure of hydrogen is 1.65 atm. The amounts of gas molecules are proportional to their representations in the flasks. If valve 2 is opened first and then the rest of the valves are opened, what will the pressure be after the first valve is opened and after they all are opened? Assume the connections have negligible volume.

 Knowledge-lite

*Text Input*

The scenario describes the application of Dalton's law.

Three 1-liter flasks are connected to a 3-liter flask by valves. The 3-liter flask is evacuated to start and the entire system is at 585 K. The first flask contains oxygen, the second hydrogen, and the third nitrogen. The pressure of hydrogen is 1.65 atm. The amounts of gas molecules are proportional to their representations in the flasks. If valve 2 is opened first and then the rest of the valves are opened, what will the pressure be after the first valve is opened and after they all are opened? Assume the connections have negligible volume.

 Knowledge-rich

*Text Input*

This problem involves the application of Dalton's law.

It involves partial pressures, which states that in a mixture of non-reacting gases, the total pressure exerted by the mixture is equal to the sum of the partial pressures of individual gases.

Three 1-liter flasks are connected to a 3-liter flask by valves. The 3-liter flask is evacuated to start and the entire system is at 585 K. The first flask contains oxygen, the second hydrogen, and the third nitrogen. The pressure of hydrogen is 1.65 atm. The amounts of gas molecules are proportional to their representations in the flasks. If valve 2 is opened first and then the rest of the valves are opened, what will the pressure be after the first valve is opened and after they all are opened? Assume the connections have negligible volume.

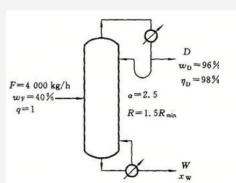
Figure 9: Examples of Different Problem Versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail

## Vision Input



Subject: Chemistry

Choices:

- A. 23.55  
 B. 45.87  
 C. 15.67  
 D. 20.77

Knowledge-free

## Text Input

As shown in the figure, the phenyl and toluene mixture is separated in a normal -pressure distillation tower. The top of the tower is used with a full condensate, bubble back, and indirect steam at the bottom of the tower. The amount of raw material treatment is 4000 kg/h, and the group is made into it\n0.4 (the mass score of benzene, the same below), the top distillation of the tower is required to become 0.96, and the recovery rate of benzene is not less than 98%. The relative volatility of the known system  $\alpha = 2.5$ , the actual return ratio is 1.5 times the minimum return ratio. The quality of the Moore of benzene and toluene is 78.11g/mol and 92.13g/mol, respectively. Try to find the amount of product top product D = \_\_\_ \$(kmol/h)\$

Knowledge-lite

## Text Input

This question involves the reaction between benzene and methanol, resulting in methylbenzene.

As shown in the figure, the phenyl and toluene mixture is separated in a normal -pressure distillation tower. The top of the tower is used with a full condensate, bubble back, and indirect steam at the bottom of the tower. The amount of raw material treatment is 4000 kg/h, and the group is made into it\n0.4 (the mass score of benzene, the same below), the top distillation of the tower is required to become 0.96, and the recovery rate of benzene is not less than 98%. The relative volatility of the known system  $\alpha = 2.5$ , the actual return ratio is 1.5 times the minimum return ratio. The quality of the Moore of benzene and toluene is 78.11g/mol and 92.13g/mol, respectively. Try to find the amount of product top product D = \_\_\_ (kmol/h)\$

Knowledge-rich

## Text Input

This question involves the reaction between benzene and methanol, resulting in methylbenzene.

The alkylation reaction between benzene and methanol, which involves the substitution of a hydrogen atom in benzene with a methyl group from methanol, results in the formation of methylbenzene.

As shown in the figure, the phenyl and toluene mixture is separated in a normal -pressure distillation tower. The top of the tower is used with a full condensate, bubble back, and indirect steam at the bottom of the tower. The amount of raw material treatment is 4000 kg/h, and the group is made into it\n0.4 (the mass score of benzene, the same below), the top distillation of the tower is required to become 0.96, and the recovery rate of benzene is not less than 98%. The relative volatility of the known system  $\alpha = 2.5$ , the actual return ratio is 1.5 times the minimum return ratio. The quality of the Moore of benzene and toluene is 78.11g/mol and 92.13g/mol, respectively. Try to find the amount of product top product D = \_\_\_ (kmol/h)\$

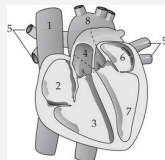
Figure 10: Examples of Different Problem Versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Chemistry


Choices:

- A. Left atrium
- B. Left ventricle
- C. Right atrium
- D. Right ventricle

 Knowledge-free

*Text Input*


In the figure, the blood enters the heart through the vena cava (\$1\$), passes through the right atrium and right ventricle and then goes through the pulmonary artery toward the lungs. After the lungs, the blood returns through the pulmonary vein and then passes into the left atrium and the left ventricle before leaving the heart via the aorta. Blood is pumped via heart contractions triggered by action potentials spreading through the heart muscle. If there is a sudden increase in blood in chamber \$3\$, which chamber of the heart received an increased number of action potentials?

 Knowledge-lite

*Text Input*

The relevant concept for this question is the cardiac conduction system.

In the figure, the blood enters the heart through the vena cava (\$1\$), passes through the right atrium and right ventricle and then goes through the pulmonary artery toward the lungs. After the lungs, the blood returns through the pulmonary vein and then passes into the left atrium and the left ventricle before leaving the heart via the aorta. Blood is pumped via heart contractions triggered by action potentials spreading through the heart muscle. If there is a sudden increase in blood in chamber \$3\$, which chamber of the heart received an increased number of action potentials?

 Knowledge-rich

*Text Input*

The relevant concept for this question is the cardiac conduction system.

The heart's action potentials begin at the sinoatrial (SA) node, spread through the atria to the atrioventricular (AV) node, then pass through the bundle of His and Purkinje fibers to the ventricles, ultimately triggering heart contractions.

In the figure, the blood enters the heart through the vena cava (\$1\$), passes through the right atrium and right ventricle and then goes through the pulmonary artery toward the lungs. After the lungs, the blood returns through the pulmonary vein and then passes into the left atrium and the left ventricle before leaving the heart via the aorta. Blood is pumped via heart contractions triggered by action potentials spreading through the heart muscle. If there is a sudden increase in blood in chamber \$3\$, which chamber of the heart received an increased number of action potentials?

Figure 11: Examples of Different Problem Versions in SCIVERSE.

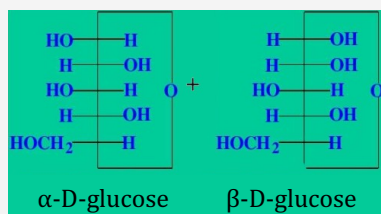


Core Question

Knowledge Cue

Knowledge Detail

*Vision Input*




Subject: Biology

Choices:


A. False

B. True

 Knowledge-free

*Text Input*


The  $\alpha$ -type and  $\beta$ -type in the same monosaccharine is a refinery. This statement is True or False?

 Knowledge-lite

*Text Input*

This question is about refinery, enantiomers and stereoisomers.

The  $\alpha$ -type and  $\beta$ -type in the same monosaccharine is a refinery. This statement is True or False?

 Knowledge-rich

*Text Input*

This question is about refinery, enantiomers and stereoisomers.

1. Refinery: A refinery is an industrial facility where raw materials like crude oil are processed and transformed into valuable products such as fuels and chemicals.
2. Enantiomers: Enantiomers are molecules that are non-superimposable mirror images of each other, similar to left and right hands.
3. Stereoisomers: Stereoisomers are compounds that have the same molecular formula and connectivity of atoms but differ in the spatial arrangement of their atoms.

The  $\alpha$ -type and  $\beta$ -type in the same monosaccharine is a refinery. This statement is True or False?

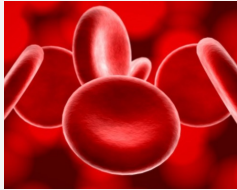
Figure 12: Examples of Different Problem Versions in SCIVERSE.

Core Question

Knowledge Cue

Knowledge Detail


*Vision Input*



Subject: Biology


Choices:

- A. 2,3-diphosphoglycerate, acidic pH
- B. Fatty acid  $\beta$ -oxidation
- C. Aerobic oxidation of sugars
- D. Glycolysis

 Knowledge-free

*Text Input*


The main energy source of mature red blood cells is:

 Knowledge-lite

*Text Input*

Mature red blood cells, also known as erythrocytes, are specialized cells in the blood responsible for transporting oxygen from the lungs to the tissues and carbon dioxide from the tissues back to the lungs.

The main energy source of mature red blood cells is:

 Knowledge-rich

*Text Input*

Mature red blood cells, also known as erythrocytes, are specialized cells in the blood responsible for transporting oxygen from the lungs to the tissues and carbon dioxide from the tissues back to the lungs.

Mature red blood cells (RBCs) lack mitochondria, so they cannot rely on aerobic oxidation of sugars (which occurs in mitochondria).

The main energy source of mature red blood cells is:

Figure 13: Examples of Different Problem Versions in SCIVERSE.